Alexandra I. Cristea · Erin Walker ·
Yu Lu · Olga C. Santos · Seiji Isotani (Eds.)

Communications in Computer and Information Science 2591

Artificial Intelligence in Education

Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED

26th International Conference, AIED 2025
Palermo, Italy, July 22–26, 2025
Proceedings, Part II

Part 2

Springer

AIED PALERMO 2025

Communications in Computer and Information Science 2591

Series Editors

Gang Li, *School of Information Technology, Deakin University, Burwood, VIC, Australia*
Joaquim Filipe, *Polytechnic Institute of Setúbal, Setúbal, Portugal*
Zhiwei Xu, *Chinese Academy of Sciences, Beijing, China*

Rationale
The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics
The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors
Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process
The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing
CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start
To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.

Alexandra I. Cristea · Erin Walker · Yu Lu ·
Olga C. Santos · Seiji Isotani
Editors

Artificial Intelligence in Education

Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED

26th International Conference, AIED 2025
Palermo, Italy, July 22–26, 2025
Proceedings, Part II

Springer

Editors
Alexandra I. Cristea
Durham University
Durham, UK

Yu Lu
Beijing Normal University
Beijing, China

Seiji Isotani
University of São Paulo
São Carlos, Brazil

Erin Walker
University of Pittsburgh
Pittsburgh, PA, USA

Olga C. Santos
UNED
Madrid, Spain

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-99263-6 ISBN 978-3-031-99264-3 (eBook)
https://doi.org/10.1007/978-3-031-99264-3

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

Welcome to the 26th International Conference on Artificial Intelligence in Education, taking place from July 22–26, 2025 in the vibrant city of Palermo, Italy. AIED 2025 is ranked A in ICORE[1], in the top 14.92% of all 784 ranked venues, and clearly growing and attracting more submissions, this year with over 800 initial submissions. Thus, as we celebrate over a quarter of a century of advancing the science and engineering of intelligent human-technology ecosystems that support learning, this edition marked a significant evolution in our conference structure, with the introduction of a dedicated track-based system. This new format was designed to capture the growth of the area, as well as the dynamic and multifaceted nature of research in educational AI, reflecting the rapid pace and expanding impact of the field.

The theme for AIED 2025 — "AI as a Catalyst for Inclusive, Personalised, and Ethical Education: Empowering Teachers and Students for an Equitable Future" — embodies our vision for a transformative future in education. Artificial Intelligence is progressing at an unprecedented rate, with generative AI and related technologies spearheading this transformative shift. As these innovations reshape both teaching and learning processes, they unlock novel opportunities for creating personalised and dynamic learning experiences. Even as traditional AI techniques continue to evolve and enrich our understanding, true change lies in leveraging these advancements to provide educators with powerful tools for deeper insights, tailored instruction, and increased student engagement.

At the heart of this transformation is the commitment to equity and ethical practice. By harnessing AI's capacity to personalise learning experiences, we can address the diverse needs of students and reduce educational disparities, especially among underserved and marginalised communities. Equally important is the need to integrate AI responsibly, ensuring that fairness, transparency, and privacy remain paramount in our educational practices. This dual commitment—to innovation and to ethical inclusion—was the cornerstone of our conference theme.

The conference program comprised seven thematic tracks:

- **Track 1 – AIED Architectures and Tools:** This track highlighted innovative engineering advances, from multi-agent systems and tangible interfaces to scalable, domain-specific applications across a variety of educational contexts.
- **Track 2 – Machine Learning and Generative AI:** Emphasising data-driven techniques and generative models, this track explored novel algorithms and computational approaches that enhance predictive and adaptive functionalities in educational environments.
- **Track 3 – Learning, Teaching, and Pedagogy:** Focusing on empirical and theoretical work, this track examined the impact of AI on teaching practices, learner engagement, and the evolution of both formal and informal learning experiences.

[1] https://portal.core.edu.au/conf-ranks.

- **Track 4 – Human-Centred Design and Design-Based Research:** This track addressed the design principles and interactive strategies that place educators and learners at the centre, fostering collaborative human-AI partnerships to promote agency and effective decision-making.
- **Track 5 – Teaching AI:** Dedicated to advancing AI literacy, the research in this track explored methods and technologies that demystify AI, empowering individuals to understand, utilise, and critically evaluate its capabilities.
- **Track 6 – Ethics, Equity, and AIED in Society:** This track delved into the societal dimensions of AI in education, with an emphasis on fairness, transparency, privacy, and strategies to mitigate educational disparities.
- **Track 7 – Theoretical Aspects of AIED and AI-Based Modeling for Education:** This track emphasised conceptual and theoretical contributions providing deeper insights into cognitive, motivational, and socio-cognitive processes, underpinning rigorous models of learning and teaching.

AIED 2025 attracted an exceptionally broad range of submissions. We received over 800 submissions across all tracks, of which 711 met the eligibility criteria and were accepted for review. From the full paper submissions, 130 were accepted as full papers, and an additional 129 were accepted as short papers. The acceptance rates for both full and short papers were 18%, respectively.

Every paper and presentation at AIED 2025 underwent a rigorous, double-blind peer-review process designed to ensure the highest scholarly standards and fair evaluation, with an average of three reviews per submission. This meticulous process—drawing on the expertise of leading researchers, seasoned reviewers, and senior committee members—culminated in an outstanding program that spanned the full spectrum of AIED research and practice.

Beyond the scholarly sessions and paper presentations, which are published in the LNCS volumes of the conference, AIED 2025 offered a rich program of keynote addresses, panel discussions, interactive demonstrations, and specialised sessions which are compiled in these CCIS volumes. The specialised tracks included **Late Breaking Results (LBR)**, **Doctoral Consortium**, **BlueSky**, **WideAIED**, and **Practitioners, Industry, and Policy (PIP)**. Furthermore, 14 **workshops** and 2 **tutorials** were held during AIED 2025.

This year had both a general AIED panel as well as a joint panel with sister conferences, Educational Data Mining (EDM 2025) and ACM Learning @ Scale (L@S 2025), which were co-located in Palermo with AIED 2025. Attendees had the opportunity to participate in the Doctoral Consortium, Workshops and Tutorials, Late-Breaking Results sessions, and dedicated tracks for practitioners, industry, and policy-makers. Notably, the BlueSky track invited visionary reflections on past progress and future trajectories in AIED, while the WideAIED track amplified perspectives from under-represented regions and communities. Additionally, the conference offered presentations on the latest publications in IJAIED2, the journal associated with the conference. As a member of the IAALDE[2] alliance, AIED 2025 hosted, last but not least, presentations of the best papers from other IAALDE conferences (in particular, ICCE 2024 and LAK 2025).

[2] https://alliancelss.com/.

This conference would not have been possible without the unwavering commitment of the International Artificial Intelligence in Education Society, the Organising Committee, the Program and Senior Program Committees, and the sponsors and partners who support our mission. Their passion and dedication to advancing intelligent educational systems is the foundation of AIED's continued success and innovation.

We invite you to engage wholeheartedly in the exchange of ideas, to challenge conventional paradigms, and to forge new collaborations that will shape the future of education. Together, we will explore how artificial intelligence can act as a true catalyst for creating more inclusive, personalised, and ethical learning environments for teachers and students around the world.

Welcome to AIED 2025 — where the future of education is actively being reimagined, one innovative idea at a time.

July 2025

Alexandra I. Cristea
Erin Walker
Yu Lu
Olga C. Santos
Seiji Isotani

Organization

Conference General Co-chairs

Olga C. Santos	UNED, Spain
Seiji Isotani	University of São Paulo, Brazil

Program Co-chairs

Alexandra I. Cristea	Durham University, UK
Erin Walker	Arizona State University, USA
Yu Lu	Beijing Normal University, China

Local Organising Co-chairs

Davide Taibi	Institute for Educational Technology, National Research Council, Italy
Giusuè Lo Bosco	University of Palermo, Italy

Doctoral Consortium Co-chairs

Zitao Liu	Jinan University, China
Carla Limongelli	University of Roma Tre, Italy
Sergey Sosnovsky	Utrecht University, Netherlands

Workshop and Tutorials Co-chairs

Gabriella Casalino	University of Bari, Italy
Tanja Mitrovic	University of Canterbury, New Zealand
Sebastiane Lalle	Sorbonne University, France

Practitioners, Industry, and Policy Co-chairs

Sreecharan Sankaranarayanan	Amazon, USA
Penghe Chen	Beijing Normal University, China
Emmanuel G. Blanchard	Le Mans University, France
Christian M. Stracke	University of Bonn, Germany

Posters and Late Breaking Results Co-chairs

Jingyun Wang	Durham University, UK
Ilona Buchem	Berliner Hochschule für Technik, Germany

AIED Mentoring Fellowship Co-chairs

Vania Dimitrova	University of Leeds, UK
Ig Ibert Bittencourt	Universidade Federal de Alagoas, Brazil

Panel Co-chairs

Ryan Baker	University of Pennsylvania, USA
Marta Cimitile	UnitelmaSapienza, Italy

Blue Sky Co-chairs

Stamos Katsigiannis	Durham University, UK
Bruce M. McLaren	Carnegie Mellon University, USA

WideAIED Co-chairs

Isabel Hilliger	Pontificia Universidad Católica, Chile
Giovanni Fulantelli	Institute for Educational Technology, National Research Council, Italy
Ilaria Torre	University of Genoa, Italy

Diversity and Inclusion Co-chairs

Valery Psyché	TÉLUQ University, Canada
Kaska Porayska-Pomsta	UCL Institute of Education, UK

Interactive Events Co-chairs

Leonardo B. Marques	Federal University of Alagoas, Brazil
Emmanuel G. Blanchard	LIUM, Le Mans Université, France

Proceedings Co-chairs

Paraskevi Topali	Radboud University, Netherlands
Rafael D. Araújo	Universidade Federal de Uberlândia, Brazil
Elaine Harada	Universidade Federal do Amazonas, Brazil

International Journal of AIED Chair

Conrad Borchers	Human-Computer Interaction Institute, USA

Virtual Experiences Co-chairs

Teng Guo	Jinan University, China
Salvatore Calderaro	University of Palermo, Italy

Publicity Co-chairs

Miguel Portaz	UNED, Spain
Xinyuan Hao	University College London, UK
Mariella Farella	Institute for Educational Technology, National Research Council, Italy

Volunteer Co-chairs

Lele Sha	Monash University, Australia
Domenico Amato	University of Palermo, Italy

Sponsorship Chair

Pietro Paolo Corso	University of Palermo, Italy

Scholarship Chair

Michele Tumminello	University of Palermo, Italy

Website Chair

Andrea Nelson Mauro	Institute for Educational Technology, National Research Council, Italy

Research Track Junior Advisory Board

Conrad Borchers	Human-Computer Interaction Institute, USA
Amine Hatun Ataş	Galatasaray University, Turkey
Krishna Chaitanya Rao Kathala	University of Massachusetts, USA
Guher Gorgun	IPN – Leibniz Institute for Science and Mathematics Education, Germany
Isaac Alpizar-Chacon	Utrecht University, Netherlands

BlueSky Program Committee

Shrikant Malviya	IIT, Patna, India
Ana Serrano-Mamolar	Universidad de Burgos, Spain
Seiji Isotani	Harvard University, USA
Peter Hastings	DePaul University, USA
Kalina Yacef	University of Sydney, Australia
Ig Ibert Bittencourt	Federal University of Alagoas, Brazil
Pablo Arnau-González	Universidad de Valencia, Spain
Vincent Aleven	Carnegie Mellon University, USA

Olga Viberg	KTH Royal Institute of Technology, Sweden
Paulo Carvalho	Carnegie Mellon University, USA
Seyedahmad Rahimi	University of Florida, USA
John Stamper	Carnegie Mellon University, USA
Donald Sturgeon	Durham University, UK
Sunčica Hadžidedić	Durham University, UK
Ivon Arroyo	University of Massachusetts Amherst, USA
Rafael Ferreira Mello	CESAR School, Brazil
Jelena Jovanovic	University of Belgrade, Serbia
Mahboobeh Mehrvarz	Carnegie Mellon University, USA
Benedict du Boulay	University of Sussex, UK
Ryan Baker	University of Pennsylvania, USA

WideAIED Program Committee

Ifeoma Adaji	University of British Columbia, Okanagan Campus, Canada
Gabriel Astudillo	Pontificia Universidad Católica de Chile, Chile
Héctor Ceballos	Tecnológico de Monterrey, Mexico
Paolo Fantozzi	LUMSA, Italy
Rafael Ferreira Mello	CESAR School, Brazil
Giovanni Fulantelli	National Research Council of Italy, Italy
Daniela Giordano	University of Catania, Italy
Isabel Hilliger	Pontificia Universidad Católica de Chile, Chile
Ivana Marenzi	L3S Research Center, Germany
Juan Andrés Talamás	Tecnológico de Monterrey, Mexico
Ilaria Torre	University of Genoa, Italy
Maomi Ueno	University of Electro-Communications, Japan
Giacomo Valente	University of L'Aquila, Italy
Mario Vallarino	University of Genoa, Italy

Practitioners, Industry, and Policy (PIP) Track

Senad Becirovic	PH NOE, Austria
Rajesh Kumar Cherukuri	Amazon, USA
Erwan David	Le Mans Université, France
Annamaria Di Grassi	Università degli studi di Bari, Italy
Guozhu Ding	Guangzhou University, China
Khaoula Ferchichi	Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

Yunlong Huang	Jiangxi University of Science and Technology, China
Sébastien Iksal	Le Mans Université, France
Dora Katsamori	NCSR "Demokritos", Greece
Imran A. S. Khan	Le Mans Université, France
Zhi Liu	Central China Normal University, China
Stefania Massaro	University of Bari, Italy
Julia Maria Mönig	Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
Nurun Nahar	University of Bolton, UK
Abdelkader Ouared	Le Mans Université, France
Dimitra Pappa	NCSR Demokritos, Greece
Yang Pian	Beijing Institute of Education, China
Claudine Piau-Toffolon	Le Mans Université, France
Marjana Prifti Skenduli	Università degli Studi di Bari "Aldo Moro", Italy
Veronica Punzo	Università di Pisa, Italy
Valérie Renault	Le Mans Université, France
Elham Tajik	University of Rochester, USA
Judith Uchidiuno	Georgia Institute of Technology, USA
Tim Walther	Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
Haipeng Wan	Capital Normal University, China
Deliang Wang	University of Hong Kong, China
Qi Wang	Beijing Foreign Studies University, China
Zhifeng Wang	Central China Normal University, China
Lan Wu	Nanjing University of Information Science & Technology, China
Sergej Zerr	Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
Jia Zhu	Zhejiang Normal University, China

Late Breaking Results (LBR) Track

Senior Program Committee Members

Adam Wynn	Durham University, UK
Daner Sun	Education University of Hong Kong, China
Emily Evermann	TU Berlin, Germany
Georges Arnaud Kouayim Bonga	Berliner Hochschule für Technik, Germany
Mandana Sutter	TU Berlin, Germany
Yi Shang	Durham University, UK
Zhongtian Sun	University of Kent, UK

Program Committee Members

Aditi Singh	Cleveland State University, USA
Aditya Joshi	UNSW, Australia
Ajanie Karunanayake	University of Technology Sydney, Australia
Allison Poh	University of Massachusetts Amherst, USA
Almazbek Beishenaliev	Regional Institute of Central Asia, Kyrgyzstan
Álvaro Sobrinho	Federal University of the Agreste of Pernambuco, Brazil
Andres Felipe Zambrano	University of Pennsylvania, USA
Andy Smith	North Carolina State University, USA
Ange Adrienne Nyamen Tato	Université Laval, Canada
Angelo Rega	Pegaso Telematic University, Italy
Anita Dąbrowicz-Tlałka	Gdańsk University of Technology, Poland
Anna Korchak	Higher School of Economics, Moscow, Russia
Anna Rafferty	Carleton College, USA
Anuradha Singh	Banaras Hindu University, India
Atsushi Shimada	Kyushu University, Japan
Bailing Lyu	University of Utah, USA
Batuhan Sayis	Universitat Pompeu Fabra/Universidad de Granada, Spain
Beate Grawemeyer	Coventry University, UK
Ben Nye	University of Southern California, USA
Boxuan Ma	Kyushu University, Japan
Bradford Mott	North Carolina State University, USA
Candace Walkington	Southern Methodist University, USA
Chaewon Kim	Florida State University, USA
Changhao Liang	Kyoto University, Japan
Chaohua Ou	Georgia Institute of Technology, USA
Cheng Tang	Kyushu University, Japan
Chenjia Zhu	University of Hong Kong, China
Christian M. Stracke	University of Bonn, Germany
Chu Hui-Chun	東吳大學, Taiwan
Chung Yiu Richard Yeung	Education University of Hong Kong, China
Cunling Bian	Ocean University of China, China
Dan Sun	Hangzhou Normal University, China
Dan R. Kohen-Vacs	Holon Institute of Technology, Israel
Daniel Alberto Perez Perez	Durham University, UK
Deliang Wang	Beijing Normal University, China
Di Zhang	Zhejiang Normal University, China
Diana-Laura Borza	Babeş-Bolyai University, Romania
Duc Nguyen	Ho Chi Minh University of Technology, Viet Nam

Eduardo Oliveira	University of Melbourne, Australia
Ehsan Latif	University of Georgia, USA
Emanuel Queiroga	Federal University of Pelotas, Brazil
Emmanuel G. Blanchard	LIUM, Le Mans Université, France
Fahim Ahmed	University of Dhaka, Bangladesh
Fatiha Sadat	UQAM, Canada
Federica Illuzzi	Università degli Studi di Bari Aldo Moro, Italy
Franziska Bickel	North Carolina State University, USA
Gabriel Astudillo	Pontificia Universidad Católica de Chile, Chile
German Capdehourat	Ceibal, Uruguay
Gila Kurtz	Holon Institute of Technology, Israel
Gongqi Lin	Victoria University, Australia
Hai Li	University of Florida, USA
Hannes Rosenbusch	University of Amsterdam, Netherlands
Hanqi Li	New York University, USA
Hao Tang	City University of New York, USA
Hideaki Yoshida	Morioka University, Japan
Hiu Ching Hung	Friedrich-Alexander University of Erlangen-Nuremberg
Hongming Li	University of Florida, USA
Huiying Cai	Jiangnan University, China
Huiyong Li	Kyushu University, Japan
Ifeoma Adaji	University of British Columbia, Okanagan Campus, Canada
Isabela Drămnesc	West University of Timisoara, Romania
Jacek Stańdo	Technical University of Lodz, Poland
Jarne Thys	Hasselt University, Flanders Make, Belgium
Jaroslaw A. Chudziak	Warsaw University of Technology, Poland
Jasbir Singh	University of Otago, New Zealand
Jennifer Kleiman	University of Georgia, USA
Jiahui Wu	Beijing University of Posts and Telecommunication, China
Jianwei Li	Beijing University of Posts and Telecommunications, China
Jiayu Zheng	Johns Hopkins University, USA
Jie Gao	McGill University, Canada
Jing Zhang	Dalian University of Technology, China
Jingxi Liu	Education University of Hong Kong, China
John Hollander	Arkansas State University, USA
Jonathan Kummerfeld	University of Sydney, Australia
Junyoung Lee	Nanyang Technological University, Singapore
Kevin Huang	WestEd, USA

Koichi Yamashita	Tokoha University, Japan
Krishna Chaitanya Rao Kathala	University of Massachusetts Amherst, USA
Leisi Pei	Education University of Hong Kong, China
Lena Hollenstein	St. Gallen University of Teacher Education, Switzerland
Liang-Yi Li	National Taiwan Normal University, Taiwan
Liangyu Chen	East China Normal University, China
Ling Li	Zhejiang Normal University, China
Lingyun Huang	Education University of Hong Kong, China
Liu Tianrui	University of Sydney, Australia
Luca Marconi	University of Milano-Bicocca, Italy
Luis P. Prieto	Universidad de Valladolid, Spain
Machi Shimmei	Tohoku University, Japan
Manikandan Ravikiran	Hitachi India Pvt. Ltd., India
Manpreet Singh	Education University of Hong Kong, China
Maomi Ueno	University of Electro-Communications, Japan
Margarida Romero	Université Côte d'Azur, France
Maria Marina	AIRI, Skoltech, Russia
Mariana Silva	University of Illinois Urbana-Champaign, USA
Marius Vogt	St. Gallen University of Teacher Education, Switzerland
Martin Ruskov	University of Milan, Italy
Matthieu Branthôme	Université de Bretagne Occidentale, France
Michael Liut	University of Toronto Mississauga, Canada
Michael Neumann	Hochschule Hannover, Germany
Min Lan	Zhejiang Normal University, China
Min Lu	Akita University, Japan
Min-Chi Chiu	嶺東科技大學資訊管理系, Taiwan
Mohammed Seyam	Virginia Tech, USA
Morgan Lee	Worcester Polytechnic Institute, USA
Mourya Teja Kunuku	Kennesaw State University, USA
Naiming Liu	Rice University, USA
Nasrin Dehbozorgi	Kennesaw State University, USA
Nghia Duong-Trung	German Research Centre for Artificial Intelligence, Germany
Nina Zhou	A*STAR, Singapore
Noshin Tahsin	Kishoreganj University, Bangladesh
Nuodi Zhang	Florida State University, USA
Ottavio G. Rizzo	Università degli Studi di Milano, Italy
Pascal Muam Mah	AGH University of Krakow, Poland
Pei-Chen Wu	National Taichung University of Education, Taiwan

R. Yamamoto Ravenor	Ochanomizu University, Japan
Rezaul Tutul	Berliner Hochschule für Technik, Germany
Rohit Saluja	IIT Mandi, India
Rositsa V. Ivanova	University of St. Gallen, Switzerland
Ruiwei Xiao	Carnegie Mellon University, USA
Ryan Zhang	Johns Hopkins University, Canada
Sabine Seufert	University of St.Gallen, Switzerland
Sabrina Eimler	Hochschule Ruhr West, Germany
Shan Zhang	University of Florida, USA
Shen Ba	Education University of Hong Kong, China
Shinobu Hasegawa	Japan Advanced Institute of Science and Technology, Japan
Shotaro Naganuma	Kyushu University, Japan
Siyuan Liu	Nanyang Technological University, Singapore
Songhee Han	Florida State University, USA
Soujatya Sarkar	Indian Institute of Technology, Mandi, India
Susan Beudt	Deutsches Forschungszentrum für Künstliche Intelligenz, Germany
Tadachika Ozono	Nagoya Institute of Technology, Japan
Uwe Maier	University of Education Schwäbisch Gmünd, Germany
Vasile Rus	University of Memphis, USA
Victor Sotelo	UNICAMP, Brazil
Vijay Prakash	Indian Institute of Technology Bombay, India
Wangda Zhu	Cornell University, USA
Xiangyu Tan	Shanghai Open University, China
Xianwen Zheng	Japan Advanced Institute of Science and Technology, China
Xinying Hou	University of Michigan, USA
Xuansheng Wu	University of Georgia, China
Yael Feldman-Maggor	KTH Royal Institute of Technology, Sweden
Yan Zhou	Zhejiang Normal University, China
Yannis Dimitriadis	University of Valladolid, Spain
Yaping Xu	Beijing Normal University, China
Yasuhiro Noguchi	Shizuoka University, Japan
Yiling Dai	Kyoto University, Japan
Yin Nicole Yang	Education University of Hong Kong, China
Yo Ehara	Tokyo Gakugei University, Japan
Yongnam Kim	Seoul National University, South Korea
Yoshimitsu Miyazawa	National Center for University Entrance Examinations, Japan
Yu Xiong	Chongqing University of Posts and Telecommunications, China

Yu-Kai Huang	Carnegie Mellon University, USA
Zechu Feng	University of Hong Kong, China
Zhi Liu	Central China Normal University/National Engineering Research Center for E-Learning, China
Zhizi Zheng	Education University of Hong Kong, China
Zifeng Liu	University of Florida, USA
Ziwei Wang	University of Sydney, Australia
Żywilla Fechner	Lodz University of Technology, Poland
昂轩 陈	北京大学, China

Doctoral Consortium Mentors

Andrew Lan	University of Massachusetts Amherst, USA
Bita Akram	North Carolina State University, USA
Bruce M. McLaren	Carnegie Mellon University, USA
Carla Limongelli	Roma Tre University, Italy
Carrie Demmans Epp	University of Alberta, Canada
Christian M. Stracke	University of Bonn, Germany
Gautam Biswas	Vanderbilt University, USA
Giora Alexandron	Weizmann Institute of Science, Israel
Ig Ibert Bittencourt	Federal University of Alagoas, Brazil
Jionghao Lin	University of Hong Kong, China
John Stamper	Carnegie Mellon University, USA
Judy Kay	University of Sydney, Australia
Kasia Muldner	Carleton University, Canada
Lauri Malmi	Aalto University, Finland
Luc Paquette	University of Illinois Urbana-Champaign, USA
Maria Mercedes Rodrigo	Ateneo de Manila University, Philippines
Michael Yudelson	Carnegie Mellon University, USA
Miguel L. Bote-Lorenzo	Universidad de Valladolid, Spain
Mingyu Feng	WestEd, USA
Neil Heffernan	Worcester Polytechnic Institute, USA
Penghe Chen	Beijing Normal University, China
Peter Brusilovsky	University of Pittsburgh, USA
Rawad Hammad	University of East London, UK
Roger Azevedo	University of Central Florida, USA
Ryan Baker	University of Pennsylvania, USA
Seiji Isotani	University of São Paulo, Brazil
Sergey Sosnovsky	Utrecht University, Netherlands
Sharon Hsiao	Santa Clara University, USA

Teng Guo	Jinan University, China
Danielle McNamara	Arizona State University, USA
Thomas Breideband	University of Colorado Boulder, USA
Tobias Ley	Danube University Krems, Austria
Valery Psyché	Université TÉLUQ, Canada
Zitao Liu	Jinan University, China

International Artificial Intelligence in Education (IAIED) Society

IAIED Management Board

President

Olga C. Santos	UNED, Spain

President-Elect

Seiji Isotani	University of São Paulo, Brazil

Secretary

Vania Dimitrova	University of Leeds, UK

Journal Editors

Vincent Aleven	Carnegie Mellon University, USA
Cristina Conati	University of British Columbia
Marcus Specht	Delft University of Technology

Finance Chair

Benedict du Boulay	University of Sussex, UK

Membership Chair

Benjamin D. Nye	University of Southern California, USA

IAIED Officers

Conrad Borchers	Carnegie Mellon University, USA
Xinyuan Hao	University College London, UK
Jionghao Lin	University of Hong Kong, China
Miguel Portaz	UNED, Spain

IAIED Executive Committee

Akihiro Kashihara	University of Electro-Communications, Japan
Amruth Kumar	Ramapo College of New Jersey, USA
Christothea Herodotou	Open University, UK
Jeanine A. Defalco	CCDC-STTC, USA
Judith Masthoff	Utrecht University, Netherlands
Maria Mercedes T. Rodrigo	Ateneo de Manila University, Philippines
Ning Wang	University of Southern California, USA
Olga C. Santos	UNED, Spain
Rawad Hammad	University of East London, UK
Zitao Liu	Jinan University, China
Bruce M. McLaren	Carnegie Mellon University, USA
Cristina Conati	University of British Columbia, Canada
Diego Zapata-Rivera	Educational Testing Service, Princeton, USA
Erin Walker	University of Pittsburgh, USA
Seiji Isotani	University of São Paulo, Brazil
Tanja Mitrovic	University of Canterbury, New Zealand
Noboru Matsuda	North Carolina State University, USA
Min Chi	North Carolina State University, USA
Alexandra I. Cristea	Durham University, UK
Neil Heffernan	Worcester Polytechnic Institute, USA
Andrew M. Olney	University of Memphis, USA
Irene-Angelica Chounta	University of Duisburg-Essen, Germany
Beverly Park Woolf	University of Massachusetts, USA
Ig Ibert Bittencourt	Universidade Federal de Alagoas, Brazil

International Alliance to Advance Learning in the Digital Era (IAALDE)

The IAIED Society is a member of the International Alliance to Advance Learning in the Digital Era (IAALDE) and belongs to its board. IAIED is currently represented at IAALDE by Olga C. Santos, as president of IAIED.

AIED 2025 Sponsors

Platinum



Gold



Silver



Bronze







Best Paper Award Sponsorship



Interactive Events Track at AIED 2025

Emmanuel G. Blanchard[1] and Leonardo B. Marques[2]

[1] Université du Mans, France
emmanuel.blanchard@univ-lemans.fr
[2] Federal University of Alagoas, Brazil
leonardo.marques@cedu.ufal.br

Presentation

The Interactive Events (IE) Track at AIED 2025 was designed to offer a dynamic platform for showcasing innovations at the intersection of education, interactive systems, and artificial intelligence. It welcomed researchers, practitioners, and industry professionals from both within and beyond the AIED community to present interactive systems and demonstrations that address a wide range of educational challenges, with the goal of fostering collaboration and advancing the field. Authors of accepted papers or posters in the main conference or other AIED tracks were particularly encouraged to participate, as the IE Track provided an additional opportunity to showcase their systems.

Contributions

IE Track contributions had to present interactive systems operating across various platforms, such as standalone applications, web-based tools, mobile devices, and more. Below is a non-exhaustive list of possible contributions:

- virtual/augmented reality environments for learning
- adaptive educational systems
- serious/educational games
- intelligent tutoring systems
- agent-based learning environments
- data-driven and/or analytic-based systems for education
- LLM-based educational technologies
- authoring tools for AIED systems
- any other interactive tool relevant to the AIED community

Contributions were required to focus on systems that had already been implemented, and to demonstrate functional features that clearly illustrated the primary educational benefits described in the submission. A selection of them was invited to be featured on the AIED Society's showcase page: https://iaied.org/showcase.

AIED 2025 Keynotes

How Generative AI Has Changed Everything and Nothing in Education

Kristen DiCerbo

Chief Learning Officer at Khan Academy

Abstract. What changes in a world with generative AI in education and what remains the same? This talk will examine this question from multiple perspectives, addressing both how students learn and how we know if they have learned something. There are fundamental, well-replicated findings about how students learn that new technology does not change. However, new technology can change the emphasis on what knowledge, skills, and attributes we deem important for students to learn. Generative AI also changes the capabilities we have at our disposal to support learning. Does it help us address some of the seemingly intractable challenges that education technology has faced in learning and assessment? Or do those challenges remain unsolved? What new challenges does generative AI introduce? I will explore all of this using the lens of lessons learned in the development, implementation, and evaluation of Khanmigo, an AI-powered tutor for students and assistant for teachers.

Bio. Dr. Kristen DiCerbo is the Chief Learning Officer at Khan Academy, where she leads the content, assessment, design, product management, and community support teams. Time magazine named her one of the top 100 people influencing the future of AI in 2024. Dr. DiCerbo's career has focused on embedding insights from education research into digital learning and assessment experiences. Prior to her role at Khan Academy, she was Vice-President of Learning Research and Design at Pearson, served as a research scientist supporting the Cisco Networking Academies, and worked as a school psychologist. Kristen has a Ph.D. in Educational Psychology from Arizona State University.

The Technopolitics of AIED: A Developing Country's Perspective

Maria Mercedes T. Rodrigo

Ateneo de Manila University

Abstract. In the IJAIED opinion piece "Is the AIED Conundrum a First-World Problem?", I argued that educational systems in developing countries such as the Philippines could greatly benefit from AI-based teaching and learning because such technologies can augment these resource-poor environments and provide students with consistent quality of instruction. The rise of generative AI and the increasing availability of extended reality applications and technologies has increased this potential for benefit. However, equity continues to be a problem. In this keynote, I will discuss the GenAI and XR developments in the Asia-Pacific, showing how these boost learning and the learning experience. I will note how innovation and use differs among developed vs. developing nations. I will discuss challenges to sustainability and scaling, e.g. a lack of hardware, software, connectivity, and teacher training as well as an inability to keep pace with changing software authoring tools' pricing models. Finally, I will discuss the implications of LLM diplomacy on developing countries. The launch of DeepSeek and similar open, free applications provides entry points for AI participation to those of us who are resource-constrained. However, we are also aware that these platforms are values-laden, and that using them for education specifically gives their creators influence over students' and teachers' values and world views. Developing nations have to balance economic and political pressures against opportunities for technological progress, deciding what tradeoffs they are willing to make in order to expand their participation in AIED.

Bio. Maria Mercedes (Didith) T. Rodrigo is a professor at the Department of Information Systems and Computer Science. Her research interests include learning analytics, artificial intelligence in education, technology in education, and educational games. She is the head of the Ateneo Laboratory for the Learning Sciences and co-lead of the Ateneo Virtual, Augmented, and Mixed Reality Laboratory. Her research work focuses on artificial intelligence in education, learning analytics, and educational games. In collaboration with colleagues and students, Dr. Rodrigo's publications focus on artificial intelligence in education, learning analytics, and educational games. She serves on the editorial boards of several high-impact journals including Artificial Intelligence in Education, the Internet and Higher Education, and Research and Practice in Technology-Enhanced Learning. Dr. Rodrigo is on the Executive Committee of the Artificial Intelligence Education Society. In 2021, Dr. Rodrigo received the Distinguished Researcher Award from the Asia-Pacific Society for Computers in Education (APSCE). She is President of APSCE. Outside of her academic work, she writes books for children. She won the National Children's Book Award Kids' Choice Award for Made Perfect in Weakness in 2016. In

2018, she won the 12th Cardinal Sin Catholic Book Awards for Cave Dweller. Her book, Blueplate, won the 2024 National Children's Book Award Kids' Choice Award and was a finalist for the Cardinal Sin Catholic Book Awards.

Beyond the Algorithm: Participatory approaches to the use of AI to support inclusive and ethical education

Francesca Pozzi

Director of CNR – Institute for Educational Technology

Abstract. The evolution of technology has always gone hand in hand with the evolution of learning theories. Artificial intelligence is no exception, as it emerges at the intersection of learning theories and – going beyond the algorithm in itself – is neither inherently 'good' nor 'bad', rather, it is the way we use it that can produce either positive or negative effects. The debate on AI highlights the potential of this technology in terms of innovation, while also warning us about the possible risks of a tool that does not necessarily provide an accurate reflection of reality and whose responses tend to reflect the values and norms of some communities, rather than others. The focus of this talk, therefore, is to reflect on how AI should be used in schools and universities to maximize the common good rather than exacerbating inequalities for the exclusive benefit of some minorities. Specifically, I argue that by promoting a participatory use of AI based on social constructivist approaches and leveraging AI interactions to encourage student exchange and discussion, we can foster the construction of new, shared, and equitable meanings. This approach can also help students reflect on and recognize the ethical risks embedded in AI systems. Promoting the use of AI as a sparring partner rather than as an expert that completes the task can lead to the enhancement of cognitive skills rather than their replacement. Discussing AI-generated results with others can encourage reflection on the structure and value of the output and can be used to stimulate students to do better than the algorithm. Additionally, it helps foster critical thinking about how misinformation can be created. However, to achieve this, educational systems need to be radically rethought and transformed in their organizational, structural, curricular, and methodological dimensions to return agency and centrality to learners and teachers need to be trained to shift their role and guide this process with due awareness. The alternative approach, consisting of simply shutting the door to this technology, would lead to schools that are even further than today from the external reality and to students prone to all the risks as soon as they leave the protected and artificial, although not intelligent, school environment.

Bio. Francesca Pozzi is Director of the Institute for Educational Technology of the National Research Council (ITD-CNR). She received her PhD in Languages, Cultures and Information and Communication Technologies from the University of Genoa and – since 1998 – she has been working as a researcher at ITD-CNR. Over the years she has coordinated and participated in numerous national and international projects, in which she has dealt with various aspects of Educational Technology. She is author of 180+ publications and is co-editor since 2016 of the journal 'Italian Journal of Educational

Technology (IJET)' (formerly 'Tecnologie Didattiche'), the first Italian scientific journal dedicated to this topic. She is also a member of Editorial Boards of several international journals and is annually involved in the scientific committees of 20+ international conferences. She is a member of the Teaching Board of the Doctorate in Digital Humanities at the University of Genoa and has been Chair of several scientific events and member of the organising committees of several scientific events.

Contents – Part II

Late Breaking Results

Using ControlNets with Generative AI Models to Create Illustrative Diagrams for Mathematics Word Problems 3
Candace Walkington and Robert DiBiano

A RAG-Enhanced AI Feedback for UML Education 12
Pasquale Ardimento, Mario Luca Bernardi, Marta Cimitile, and Michele Scalera

A Zero-Shot LLM Framework for Automated Assignment Grading in Introductory Higher Education Courses 19
Calvin Yeung, Jeff Yu, King Chau Cheung, Tat Wing Wong, Chun Man Chan, Kin Chi Wong, and Keisuke Fujii

From Code to Insight: How LLMs Help and Hinder Qualitative Research 28
Chaewon Kim, Fengfeng Ke, Alex Barrett, and Nuodi Zhang

Grading University Students with LLMs: Performance and Acceptance of a Canvas-Based Automation 36
Luke Korthals, Hannes Rosenbusch, Raoul Grasman, and Ingmar Visser

Short-Term Gains, Long-Term Gaps: The Impact of GenAI and Search Technologies on Retention 44
Mahir Akgun and Sacip Toker

Enhancing Teacher Support in Learning Technologies: A Human-Centered Approach with WearableLearning 53
Allison Poh, Yuqian Shi, Francisco Enrique Vicente Castro, and Ivon Arroyo

SSRLBot: Designing and Developing a Large Language Model-Based Agent Using Socially Shared Regulated Learning 61
Xiaoshan Huang, Jie Gao, and Haolun Wu

TPRS: AI-Assisted Research Topic Refinement for Distance Learners 69
Nghia Duong-Trung, Xia Wang, Rahul Rajkumar Bhoyar, Angelin Mary Jose, Silke Elisabeth Wrede, Lars van Rijn, Theresa Panse, Claudia de Witt, and Niels Pinkwart

Facilitating Teacher Coaching: Generating Insightful Comments on Classroom Instruction Through Multi-agent Reflection 77
Jinglei Yu and Yu Lu

Modeling Socially Constructed Knowledge Using Multimodal Machine Learning: A Case Study in K-12 AI Literacy Education Classroom 85
Md Biplob Hosen, Lujie Karen Chen, Sancharee Hom Chowdhury, Erin Bosarge, Na Gong, Woei Hung, Becky Tomaso, and Shenghua Zha

beSEL: A Human-Aligned LLM Teaching Assistant for Enhancing Social-Emotional Learning 93
Jing Zhang, Yazhe Niu, Xueyan Li, Peiyan Zhou, and Di Sun

Automated Essay Scoring Using Logical Structure Graph Information 102
Yoshihiro Kato

Leveraging AI for Persons with Disabilities: Investigating Their Needs and Gaps in AI-Supported Assistance 110
Rolf Feichtenbeiner, Lorenz Reichert, and Susan Beudt

AI-Powered Math Learning: Evaluating the Impact of a Personalized Tutoring Platform that Responds to Affect 118
Mingyu Feng, Natalie Brezack, Chunwei Huang, Megan Schneider, Ivon Arroyo, Beverly Woolf, and Danielle Allessio

Tell Me Who Your Students Are: GPT Can Generate Valid Multiple-Choice Questions When Students' (Mis)Understanding is Hinted 126
Machi Shimmei, Masaki Uto, Yuichiroh Matsubayashi, Kentaro Inui, Aditi Mallavarapu, and Noboru Matsuda

Leveraging Large Reasoning Models for Test Equating Without Anchor Items: A Simulation Study with O1 and DeepSeek-R1 135
Junlei Du, Qinhua Zheng, and Shuang Li

Leveraging Knowledge Profiles and Generative AI for Realistic Student Response Generation 143
Eylül Ipçi, Tanya Nazaretsky, and Tanja Käser

LLMs to Support K–12 Teachers in Culturally Relevant Pedagogy: An AI Literacy Example 152
Jiayi Wang, Ruiwei Xiao, Xinying Hou, Hanqi Li, Ying Jui Tseng, John Stamper, and Kenneth Koedinger

Comparison of Explainable Machine Learning Methods for Early Prediction of Student Performance in Programming Courses 161
Ziwei Wang, Irena Koprinska, María Martínez-Ballesteros, Alicia Troncoso, and Bryn Jeffries

ScoreCLIQ: A Dynamic LLM-Based Framework for Item Difficulty Estimation 169
Soujatya Sarkar, Manikandan Ravikiran, and Rohit Saluja

LLM Agents for Verifiable Question Generation and Grading 177
Jacob Levine, Matt West, and Mariana Silva

Computer Vision for Objects Used in Group Work: Challenges and Opportunities 185
Changsoo Jung, Sheikh Mannan, Jack Fitzgerald, and Nathaniel Blanchard

From Sketch to Understanding: Exploring LLM-Based Assessment of Student-Drawn Science Models 193
Andy Smith, Natalie Brezack, Wynnie Chan, Abby Lavine, Bradford Mott, Cathy Ringstaff, Mingyu Feng, and James Lester

Riding on the Back of a Whale: A Hackathon Framework for Introducing High School Students to Large Language Models 201
Duc Nguyen, Dong Le, Long Nguyen, Quyen Vu, Tran Le, Dung Nguyen, Nga Huynh, Huong Nguyen, Phat Tran, Dang Le, Sang Truong, Sanmi Koyejo, Cuong Le, and Tho Quan

Design and Validation of the Psychometric Properties of a Questionnaire Measuring AI Literacy Among University Students: Preliminary Results 210
Chantal Tremblay, Bruno Poellhuber, Anne-Marie Sénécal, and Normand Roy

ChatGPT's Impact on Students' Performance in Solving Tasks Requiring Critical Thinking 218
Kira Rohwer and Sabine Seufert

Intent-Based vs. LLM-Based Chatbots in MOOCs: Do Diverse Students Experience Them Differently? 226
Songhee Han, Chaewon Kim, and Hunhui Na

A GenAI-Enhanced Platform for Personalized Doctoral Progress and Well-Being Support Through Single-Case Analytics 234
Mohamed Saban, Luis P. Prieto, and Yannis Dimitriadis

AGen: Personalized Analogy Generation with Large Language Model 242
Shutong Wu, Hecong Wang, and Zhen Bai

'The World of AI': A Novel Approach to AI Literacy for First-Year Engineering Students 250
Siddharth Siddharth, Brainerd Prince, Amol Harsh, and Shreyas Ramachandran

Exploring Communication Strategies for Collaborative LLM Agents in Mathematical Problem-Solving 258
Liang Zhang, Xiaoming Zhai, Jionghao Lin, Jennifer Kleiman, Diego Zapata-Rivera, Carol Forsyth, Yang Jiang, Xiangen Hu, and Arthur C. Graesser

VETTING AI for Deeper Learning: Constraining LLMs to Encourage Student Inquiry 266
Shan Zhang, Hongming Li, Seiyon M. Lee, Noah L. Schroeder, and Anthony F. Botelho

Exploring the Feasibility of Personalized AI Feedback to Improve Children's Planning Skills 274
Somphop Sukjaitham, Garvin Brod, and Jasmin Breitwieser

Exploring AI Chatbots for Learning Mathematics: Students' Perspectives on Accuracy and Educational Value 282
Jacek Stańdo, Żywilla Fechner, Anita Dąbrowicz-Tlałka, Katarzyna Kujawska, and Magdalena M. Musielak

Making Generative AI Hallucinations Useful by Reassessing the Troublemaker Agent Strategy 289
Emmanuel G. Blanchard, Jean-Christophe Callahan, Idir Akretche, Naif Asswiel, Lucas Schmitt, Amine Kessemtini, and Imran S. A. Khan

Enhancing Collaborative Academic Writing with Generative Artificial Intelligence: Use and Effects 297
Anna Korchak, Mik Fanguy, Kseniia Adamovich, Han Zhang, Mattew Baldwin, and Jamie Costley

Exploring Students' Generative AI Usage Patterns and Knowledge Creation in Collaborative Problem Solving 305
Shotaro Naganuma, Tsubasa Minematsu, Sachika Shibukawa, Aikana Ohno, and Yukinori Wakihama

Leveraging Multi-modality and Collaborative Filtering for Supporting Automatic Scoring in Mathematics Education 313
Hai Li, Wanli Xing, Wangda Zhu, Chenglu Li, Bailing Lyu, Zifeng Liu, and Neil Heffernan

Developing Artificial Intelligence Explanations for Technology Enhanced Learning with and for Special Educational Needs Children and Teachers 321
Beate Grawemeyer and Manolis Mavrikis

SPAR-GNN: Knowledge Tracing with Behavioural Patterns and Selective LLM Feedback 328
Zhongtian Sun, Jingyun Wang, Ahmed Alamri, and Alexandra Cristea

How far are we from Automatic Grading of Handwritten Cloze Form Questions? 336
Shrey Chandola, Manikandan Ravikiran, and Rohit Saluja

ProductiveMath: A Generative-AI-Powered App to Support Productive Failure Teaching 344
Seyedahmad Rhaimi, Deniz Ercan, Ran Gao, Salah Esmaeiligoujar, Maryam Babaee, Hongming Li, Shan Zhang, Seiyon Lee, Avery Closser, and Anthony Botelho

Automated Bias Assessment in AI-Generated Educational Content Using CEAT Framework 352
Jingyang Peng, Wenyuan Shen, Jiarui Rao, and Jionghao Lin

Mining Student Activity Sequences in a Large-Scale Online Programming Course 359
Ziwei Wang, Bryn Jeffries, and Irena Koprinska

Assessing Lecture Difficulty and Unexpected Events Using Subtle Facial Cues 366
Tadachika Ozono, Yuna Kaneko, and Masato Kikuchi

Author Index 375

Late Breaking Results

Using ControlNets with Generative AI Models to Create Illustrative Diagrams for Mathematics Word Problems

Candace Walkington[1(✉)] and Robert DiBiano[2]

[1] Southern Methodist University, Dallas, TX 75205, USA
cwalkington@smu.edu
[2] Ailectric, Baton Rouge, LA 70810, USA
robert@ailectric.com

Abstract. The generation of illustrative diagrams to accompany mathematics word problems has the potential to provide adaptive support for students who struggle with mathematics. However, typical approaches to AI-generated illustrative diagrams have not been successful, with image either being visually appealing or mathematically accurate, and not both. Here we present a novel method for generating illustrative diagrams using generative AI – ControlNets. We demonstrate how ControlNets can be used with open educational resources to create mathematically precise, visually compelling illustrative diagrams to accompany mathematics story problems.

Keywords: ControlNet · math word problem · illustration · proof-of-concept

1 Background

Math word problems (MWPs), verbal descriptions of problem situations that can be solved by applying math procedures, are challenging for many students [1–3]. Illustrative diagrams added to MWPs can be a form of adaptive support to show a problem's real-world context and mathematical relationships, which are often communicated in text alone [3–5]. Developing mathematically precise and visually compelling illustrations is expensive, and recent advances in generative AI (GenAI) programs like Stable Diffusion [6] have offered new avenues for creating artwork. However, typical GenAI do not render images with precise mathematical properties (see [7]). Indeed, GenAI's current lack of ability to supply useful visuals to mathematics learners has been repeatedly noted as a major limitation of these models [8–10]. In this conceptual proof-of-concept paper, we demonstrate a novel method for rendering illustrative diagrams that maintain mathematical precision while also being visually engaging, using ControlNets [11].

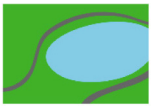
Our method is designed to particularly benefit those who produce and use open educational resources (OER) [12]. OER, which have flexible licensing and may be revised and redistributed [13], are a critical avenue for access, as they offer students, schools, and districts a free alternative for curriculum materials. GenAI has been proposed as a way

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 3–11, 2025.
https://doi.org/10.1007/978-3-031-99264-3_1

to improve the quality of OER [14, 15] and research suggests that OER texts have lower quality visuals than their major-publisher counterparts [16–18]. In contemporary OER like the middle school curriculum *Illustrative Mathematics* (IM) [19], image generation is clearly an issue. IM is a widely-used math curriculum for grades K-12 in the United States. In IM, many problems that could incorporate visuals do not have visuals. For example, the IM unit on Area and Surface Area, which should be a ripe area for rich visuals, has virtually *no images* that contain illustrative features or real-world objects. These concepts are visualized in a completely abstract manner, giving the impression of math as disconnected from the world. This is especially ironic for a curriculum that is named "Illustrative Mathematics." Other problems reference vocabulary words like "cargo ship," "pasta dough," "slug," "a green" (in golf), and other terms students may find difficult, with no visuals.

The visuals that are in IM can be low quality or misleading (Table 1). The mice image in Table 1 is misleading. None of the mice are clearly white or brown as in the problem – most are mixed or grayish. The image does not match the information given, where 25–40% of the mice would be expected to be white. Other images, like the park in Table 1, are so simple and low quality it can be difficult to decode what they are supposed to depict. The third example in Table 1 is mathematically inaccurate – the image shows 1/5 of the box being shaded for "20% more," when it should be 1/6 (i.e., the original 5/5 plus one more 1/5); this happens when artists who do not have mathematics education training draw images. In addition, the cereal box is rendered in a simplistic manner that in no way realistically evokes the look and feel of a modern cereal box.

Table 1. Existing problem text with illustrative diagrams in Illustrative Mathematics OER

Problem Type	Example Problem from *Illustrative Mathematics*	Associated Image from *Illustrative Mathematics*
Collection/Set	When two mice with brown fur breed, there is a 25% chance that each baby will have white fur. For the experiment to continue, the scientist needs at least 2 out of 5 baby mice to have white fur.	
Geometric-2D	For example, here is a scale drawing of a park where the scale is 1 cm to 90 m. The rectangle is 10 cm by 4 cm, so the actual dimensions of the park are 900 m by 360 m.	
Geometric-3D	A cereal box says that now it contains 20% more. Originally, it came with 18.5 ounces of cereal. How much cereal does the box come with now?	

1.1 Theoretical Framework

Theoretically, we frame the considerations of illustrations with respect to interest theory, the multimedia principle, and seductive details. Interest is defined as the psychological state of engaging *and* the predisposition to re-engage [20]. Higher levels of interest are associated with improved performance and learning (see [21]). *Situational interest* is an immediate and transient response to a stimulus that involves heightened affect and the allocation of attentional resources [20, 22]. For visuals, research suggests that concreteness, comprehensibility, and personal relevance can trigger interest [23]. Mayer's [24] *multimedia principle* states that words with visuals will be more understandable than words alone, because students can make connections among different ideas when using two distinct mental representations - the verbal representation and the visual representation, which occur through two channels of processing - auditory/verbal and visual/pictorial [25]. As each system is individually limited in terms of its work memory capacity, leveraging both systems is advantageous. Research has examined the effect of adding *seductive details* - interesting but unimportant information, including visuals – to tasks [26]. A meta-analysis [27] of 68 experiments found an overall negative effect of seductive details ($g = -0.41$). However, the analysis found a *positive* effect when seductive details were presented in math materials ($g = 0.42$; seductive details were often in the form of images). Thus, it is not clear math visuals will have detrimental effects due to seductiveness, particularly since the visuals we use have math-relevant features. The degree to which visuals are detrimental due to seductiveness likely depends on the type of visual used – thus we next review four types of visuals.

1.2 Literature Review

We define four types of visuals, based on [30]. *Decorative illustrations* do not relate to the math task (e.g., a picture of a student doing homework); these may be the kind of visual most in danger of adding seductive details to learning tasks. *Contextual illustrations* generally show the meaning of words in a mathematical task (e.g., a picture of a roller coaster), but do not communicate any mathematical information; these may also have elements of seductive details. *Diagrams* are precise technical mathematical figures that do not necessarily have any real-world features – they are completely abstract, and thus unlikely to add seductive details. *Illustrative diagrams* both communicate mathematical information precisely and illustrate real-world contexts and ideas in an engaging manner – this is the kind of visual we are focusing on in the present paper.

GenAI programs like Flux [31] and Stable Diffusion [32] excel at generating decorative and contextual illustrations. However, they do not perform well drawing illustrative images with mathematical properties. Conwell et al. [6] explained this difficulty as an inability to deal with logical operators – including relations (e.g., positioning one object under another or inside another), negations (e.g., understanding prompts that indicate what something is not), and numbers (e.g., drawing an exact integer number of objects). Hardison and Gómez Marchant [6] examined pre-service teachers using Adobe Firefly [33] to illustrate elementary mathematics story problems and found the GenAI unable to generate sensible images for most scenarios. Li et al. [34] present a platform for grades 3–5 called *Read Solve Create*, that uses DALL-E3 [35] to generate images. Their

research suggests that semantic similarity between the image and the story problem text reduces cognitive load. However, all images they present in their article are problematic. An image of 14 snacks accompanies a problem text describing 30 snacks, an image with a pizza cut into 10 slices accompanies a text scenario about a pizza cut into 8 slices, and a scenario about ¼ of a garden row being planted has an image with only completely-planted rows that already have mature plants. GenAI can generate non-illustrative diagrams. For example, users can ask ChatGPT [36] to use a Python package like *mathplotlib* to draw a diagram using plain descriptive language. Then, if the image is not accurate, users can ask ChatGPT in plain language to modify the diagram in various ways. However, these diagrams are not illustrative or realistic – they are generated by very simple computer rendering algorithms. In this conceptual paper, we pursue the question: *How can ControlNets be used to create illustrative diagrams for mathematics story problems that are both mathematically accurate and visually appealing*?.

2 Image Generation Method

Diffusion models, a class of GenAI models used to create digital images from textual descriptions, function by treating random noise as a noisy image, and iteratively denoising the image until it matches the text description to create new synthetic data. Stable Diffusion [31] is one of the most popular diffusion models and is free and open source [37]. We leverage the recent technical advance of ControlNets [38] which came out in 2023 for diffusion models. ControlNets are neural network structures or architectures that allow for the addition of extra conditions to diffusion models. ControlNets create a copy of an original neural network, keeping one copy locked to function as the model and the other trainable to generate new images based on the information from the locked model. ControlNets can extract outlines, depth maps, segmentation maps, human poses, as well as edges and lines, and use these to guide subsequent image generation.

ControlNets are considered revolutionary because they can instruct an image generation model to use particular aspects of an inputted image with great precision, allowing for strongly enforced spatial consistency [39]. However, given their newness, there are still many research avenues that need to be pursued, such as which type of ControlNet model works best for different image types, how to set up ControlNets to optimally preserve precise features from the original image, and how to streamline the process of generating images with ControlNets. ControlNets largely came about in computer vision as a way to generate images of humans in particular body poses; this is their primary use case. Thus, our application case is novel and has important potential to advance research in computer vision and artificial intelligence.

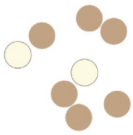
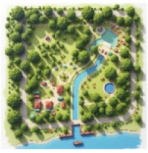
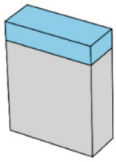
Until recently, ControlNets had only been implemented for Stable Diffusion [40]. However, a new free, open-source AI image generation engine Flux (produced by Black Forest Labs [31]) now also has ControlNets. Flux has a hybrid architecture where it can combine transformer techniques, which power contemporary LLMs and allow their understanding of language, with diffusion techniques [41]. As a result, Flux is better at understanding language and generating images based on textual language prompts - it can generate more abstract concepts inside of the image generation weights. This improved language model is especially important for more complex prompts, which may be necessary to generate some types of illustrative diagrams.

Our current approach has three steps. First, we prompt GPT-4o to use the matplotlib library in Python to render a math diagram, giving it instructions in natural language (e.g., "Draw a half of a sphere that has a circumference of 20 cm"). GPT-4o generates code and feeds it to an online version of Python. Second, the image is exported and goes into pre-processing, generating a depth map, outline map, segmentation map, etc., depending on the type of ControlNet being used. We generally find that depth maps (for 2D objects and collections) and soft edge maps (for 3D objects) are appropriate. This preprocessing is done through either DreamshaperXL [42], a fine-tuned version of Stable Diffusion, or FLUX.1-dev [43]. Third, the pre-processed image is fed into the ControlNet within Flux or DreamshaperXL, with a second prompt used to specify what illustrative features to layer on top of the diagram (e.g., "Create a mathematics textbook illustration of half an orange. Design features a spherical orange, cut in half, sharp clean cut, placed on a bright white background."). The final illustrative diagram is then output. We created one large standard prompt, and then gave special instructions for specific types of problems, letting the language model decide which apply to a given situation. For rendering 3D shapes, we instructed it to draw a solid shape rather than a translucent mesh, providing some example functions to do this. For rendering collections of objects, we had additional steps where it first drew circles with mathplotlib of where we wanted each object to appear on the coordinate plane, and then we used a ControlNet to create intermediate round semi-finished anchor objects. Next, we used the hue channel of the plot to color the circles over the intermediate objects, and then used image-to-image to convert the round objects to final objects that were no longer circular.

3 Examples

Examples are shown in Table 2. For the mice scenario, Stable Diffusion alone generates an image that is visually compelling but that does not match the problem. The mice do not clearly have either white or brown fur, and the ratio of white to brown fur is not apparent. Python draws a diagram where the mice are represented by circles, and the circles are colored in an appropriate ratio. When the ControlNet was applied, the image is mathematically precise while also being visually appealing. For the park scenario, Flux created a beautiful overhead map of the park that is roughly a rectangular shape. However, it is not able to provide the correct ratio of side lengths of 10:4. Python creates this abstract diagram easily, and then the ControlNet layers the image of a beautiful park over this accurate rectangular shape. For the final scenario about a cereal box, Flux creates a visually compelling rectangular prism cereal box, but many attempts did not yield a result where it could shade the top one-sixth of the box. Python was able to render a non-translucent diagram, which then was easy to layer the ControlNet to produce an accurate image.

Table 2. Images generated with Stable Diffusion/Flux, Python, and our ControlNet approach

Problem Type	Stable Diffusion/Flux Alone	Python via ChatGPT	ControlNet
Collection/Set *(6 brown and 2 white mice)*			
Geometric-2D *(rectangle park with 10:4 side ratio)*			
Geometric-3D *(rectangular prism cereal box with top 1/6 shaded)*			

4 Discussion and Conclusion

In this paper, we described the novel method of using ControlNets to generate mathematically precise, visually appealing illustrative diagrams. We confront what has been noted in the literature as a major limitation of GenAI [7–9] – its ability to generate meaningful mathematical diagrams for MWPs. We also offer an important potential alternative for current online learning systems which are using GenAI to give children mathematically inaccurate diagrams [35], offering a solution to make these diagrams reflect the mathematical properties of the problem while still being visually appealing. We also provide an important future consideration for OER developers. Future user research should be conducted with mathematics students to address how contextual illustrations made with ControlNets elicit interest [19], potentially add seductive details [26], and/or allow for better processing of information [23]. In addition, ethical issues with the generated images – including the degree to which images show cultural bias – should be evaluated; this is a direction of our current work. We are using multi-agent systems that evaluate generated images based on human-created rubrics for cultural bias.

An important class of diagrams that we did not include in our method that are present in *Illustrative Mathematics* and other curricula are number line diagrams – these include images like a thermometer or a ruler, with different labelled locations along a vertical or horizontal line, sometimes incorporating illustrative features. It may be that many of these illustrative diagrams can be produced by using our method for 2D geometric shapes, using a long thin rectangle, but more research is needed. More research is also needed to understand the affordances and limitations of being able to add textual and numerical labels to diagrams during the generation process; the latest version of

ChatGPT's image generator excels at adding text. Flux was also able to relatively easily incorporate the "20% more" text into the cereal box, in a way that was not possible with Stable Diffusion. However, concerns have been raised about image-generating AI utilizing human artists' work in their training sets without permission. Stable Diffusion 3 [44] (the current version) gave artists the opportunity to opt-out of being included in its training set and is thus more ethical than many of the alternatives. But any consideration of using these programs in OER [12] should meaningfully engage with the potential ethical consequences for artists.

ControlNets can generate images that would be impossible by traditional prompting approaches and learning more about how to apply them in contexts where the composition of the image and its spatial positioning and measurements need to be incredibly precise will have far-reaching implications across many different use cases. We generally found that Flux follows the prompts better than Stable Diffusion, given its underlying language model, making it our recommendation for future use. However, ControlNets are less developed for Flux, thus important future development must take place. Our techniques could be applied to technical diagrams in textbooks in other fields – like science or engineering, as well as technical diagrams used in industry. Eventually, systems could be developed where educational tasks could be entered into an app, and one or several illustrative diagrams could be generated on-the-fly for teachers or curriculum developers to select from to use in their materials, or for students to get direct assistance from.

Acknowledgments. This work was supported by the National Science Foundation under Grant DRL 2341948. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Verschaffel, L., Schukajlow, S., Star, J., et al.: Word problems in mathematics education: a survey. ZDM **52**, 1–16 (2020)
2. Xu, C., et al.: The role of mathematical language skills in arithmetic fluency and word-problem solving for first-and second-language learners. J. Educ. Psychol. **114**(3), 513 (2022)
3. Martiniello, M.: Language and the performance of English-language learners in math word problems. Harv. Educ. Rev. **78**(2), 333–368 (2008)
4. Barton, B., Neville-Barton, P.: Investigating the relationship between English language and mathematical learning. In: Proceedings of the Third Conference of the European Society for Research in Mathematics Education, pp. 1–10 (2003)
5. Wolf, M.K., Leon, S.: An investigation of the language demands in content assessments for English language learners. Educ. Assess. **14**(3–4), 139–159 (2009)
6. Stability AI. Stable Diffusion. Stability AI. https://stability.ai
7. Conwell, C., Tawiah-Quashie, R., Ullman, T.: Relations, negations, and numbers: looking for logic in generative text-to-image models. arXiv preprint arXiv:2411.17066

8. Gattupalli, S.S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., Woolf, B.P., Woolf, B.: Exploring pre-service teachers' perceptions of large language models-generated hints in online mathematics learning. In: LLM@ AIED, pp. 151–162 (2023)
9. Hardison, H.L., Gómez Marchant, C.N.: Illustrated story problems: prospective teachers' mathematically pedagogical usage of text-to-image generative artificial intelligence. In: Proceedings of the Psychology of Mathematics Education – North America conference, 7–10 November 2024 (2024)
10. Yadav, G., Tseng, Y.J., Ni, X.: Contextualizing problems to student interests at scale in intelligent tutoring system using large language models. arXiv:2306.00190 (2023)
11. Stability AI. ControlNets for Stable Diffusion 3.5 Large. Stability AI. https://stability.ai/news/sd3-5-large-controlnets. Accessed 15 Jan 2025
12. Open Educational Resources. OER Commons: Open Educational Resources. Open Educational Resources. https://oercommons.org/. Accessed 15 Jan 2025
13. Green, C.: Open licensing and open education licensing policy. In: Jhangiani, R.S., Biswas-Diener, R. (eds.) Open: The Philosophy and Practices that are Revolutionizing Education and Science, pp. 29–41 (2017). https://www.jstor.org/stable/j.ctv3t5qh3.7
14. Bozkurt, A.: Generative AI, synthetic contents, open educational resources (OER), and open educational practices (OEP): a new front in the openness landscape. Open Praxis **15**(3), 178–184 (2023). https://doi.org/10.55982/openpraxis.15.3.579
15. Tlili, A., Burgos, D.: Unleashing the power of Open Educational Practices (OEP) through Artificial Intelligence (AI): where to begin? Interact. Learn. Environ. **32**(10), 6886–6893 (2022)
16. Clinton, V., Legerski, E., Rhodes, B.: Comparing student learning from and perceptions of open and commercial textbook excerpts: a randomized experiment. Front. Educ. **4**, 110 (2019). https://doi.org/10.3389/feduc.2019.00110
17. Howard, V.J., Whitmore, C.B.: Evaluating student perceptions of open and commercial psychology textbooks. Front. Educ. **5**, 139 (2020)
18. Price, J.L.: Textbook bling: An evaluation of textbook quality and usability in open educational resources versus traditionally published textbooks. Brigham Young University (2012)
19. Illustrative Mathematics. Illustrative Mathematics, Learn Math for Life. Illustrative Mathematics. https://www.illustrativemathematics.org. Accessed 15 Jan 2025
20. Hidi, S., Renninger, K.A.: The four-phase model of interest development. Educ. Psychol. **41**(2), 111–127 (2006)
21. Renninger, K.A., Hidi, S.E.: Interest: a unique affective and cognitive motivational variable that develops. In: Advances in Motivation Science, p. 179 (2022)
22. Hidi, S.E.: A reexamination of the role of attention in learning from text. Educ. Psychol. Rev. **7**, 323–350 (1995)
23. Magner, U.I.E., Glogger, I., Renkl, A.: Which features make illustrations in multimedia learning interesting? Educ. Psychol. **36**(9), 1596–1613 (2016)
24. Mayer, R.E.: Applying the science of learning: evidence-based principles for the design of multimedia instruction. Am. Psychol. **63**(8), 760–769 (2008)
25. Mayer, R.E.: Cognitive theory of multimedia learning. Cambridge Handb. Multimedia Learn. **41**(1), 31–48 (2005)
26. Garner, R., Gillingham, M.G., White, C.S.: Effects of seductive details' on macroprocessing and microprocessing in adults and children. Cogn. Instr. **6**(1), 41–57 (1989). https://doi.org/10.1207/S1532690XCI0601_2
27. Sundararajan, N., Adesope, O.: Keep it coherent: a meta-analysis of the seductive details effect. Educ. Psychol. Rev. **32**(3), 707–734 (2020)
28. Woodward, J., et al.: Improving mathematical problem solving in grades 4 through 8. IES Practice Guide. NCEE 2012–4055. What Works Clearinghouse (2012)

29. Python Software Foundation. Python programming language. Python. https://www.python.org. Accessed 15 Jan 2025
30. Cooper, J., Sidney, P., Alibali, M.: Who benefits from diagrams and illustrations in math problems? Ability and attitudes matter. Appl. Cogn. Psychol. **32**(1), 24–38 (2018)
31. Black Forest Labs. Flux: A creative tool for design. Black Forest Labs. https://www.blackforestlabs.com/flux. Accessed 15 Jan 2025
32. Stability AI. Activating humanity's potential through generative ai. Stability AI. https://stability.ai. Accessed 15 Jan 2025
33. Adobe Inc. Adobe Firefly; Create with Adobe Firefly generative AI. Adobe. https://www.adobe.com/products/firefly.html. Accessed 15 Jan 2025
34. Li, H., Xing, W., Li, C., Zhu, W., Oh, H.: Are simpler math stories better? Automatic readability assessment of GAI-generated multimodal mathematical stories validated by engagement. Brit. J. Educ. Technol. (2024)
35. Open AI. DALL-E3. https://openai.com/index/dall-e-3/
36. Open AI. ChatGPT4. https://openai.com/index/gpt-4/
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. arXiv (2022)
38. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv, preprint arXiv:2302.05543
39. Ionio. Extensive guide to ControlNet: Controlling AI generated Images. Ionio 2023 (2023). https://www.ionio.ai/blog/extensive-guide-to-controlnet-controlling-ai-generated-images. Accessed 15 Jan 2025
40. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. arXiv, preprint arXiv:1503.03585 (2015)
41. Esser, P., et al.: Scaling rectified flow transformers for high-resolution image synthesis. arXiv. https://arxiv.org/abs/2403.03206 (2024)
42. Stability AI. DreamshaperXL. Stability AI. https://stability.ai. Accessed 15 Jan 2025
43. Black Forest Labs. FLUX.1.dev. Black Forest Labs. https://huggingface.co/black-forest-labs/FLUX.1-dev
44. Stability AI. Stable Diffusion 3. Stability AI. https://stability.ai/news/stable-diffusion-3. Accessed 15 Jan 2025



A RAG-Enhanced AI Feedback for UML Education

Pasquale Ardimento[1(✉)], Mario Luca Bernardi[2], Marta Cimitile[3], and Michele Scalera[1]

[1] Department of Informatics, University of Bari Aldo Moro, Bari, Italy
{pasquale.ardimento,michele.scalera}@uniba.it
[2] Department of Engineering, University of Sannio, Benvento, Italy
bernardi@unisannio.it
[3] Department of Law and Digital Society, UnitelmaSapienza University, Rome, Italy
marta.cimitile@unitelmasapienza.it

Abstract. Software modeling education often lacks immediate and personalized feedback, making it challenging for novice learners to comprehend modeling principles. This paper continues our ongoing work on developing an AI-driven feedback mechanism to support UML diagram construction. Building on previous efforts, we improved the system by enhancing the RAG-LLM component within the existing UML Miner plugin. The system analyzes students' modeling behavior and provides real-time, personalized guidance to support learning and improve modeling skills.

Keywords: Artificial Intelligence · UML class diagrams · Software Modeling Education · Retrieval Augmentation Generation

1 Introduction

Unified Modeling Language (UML) is a graphical language for visualizing, specifying, constructing, and documenting the artifacts of object-oriented systems. It provides a set of diagram types, such as class, sequence, and activity diagrams, that enable developers and stakeholders to visualize system structure and behavior. As such, UML serves as a valuable pedagogical tool for teaching core principles of software design and abstraction. Improving software modeling skills among students and novice modelers is a significant goal in software engineering education. Differently from existing educational UML tools such as QuickUML [4], AutoER [5], and SD4ED [1], our approach supports learners by providing immediate, personalized feedback during modeling activities. The tool monitors interactions within the Visual Paradigm environment, recording event logs that reflect user modeling behavior. These logs, along with the task requirements and a natural language description of the diagram, are analyzed by a Retrieval-Augmented Generation Large Language Model (RAG-LLM). The RAG-LLM leverages a structured knowledge base to generate context-aware, personalized

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 12–18, 2025.
https://doi.org/10.1007/978-3-031-99264-3_2

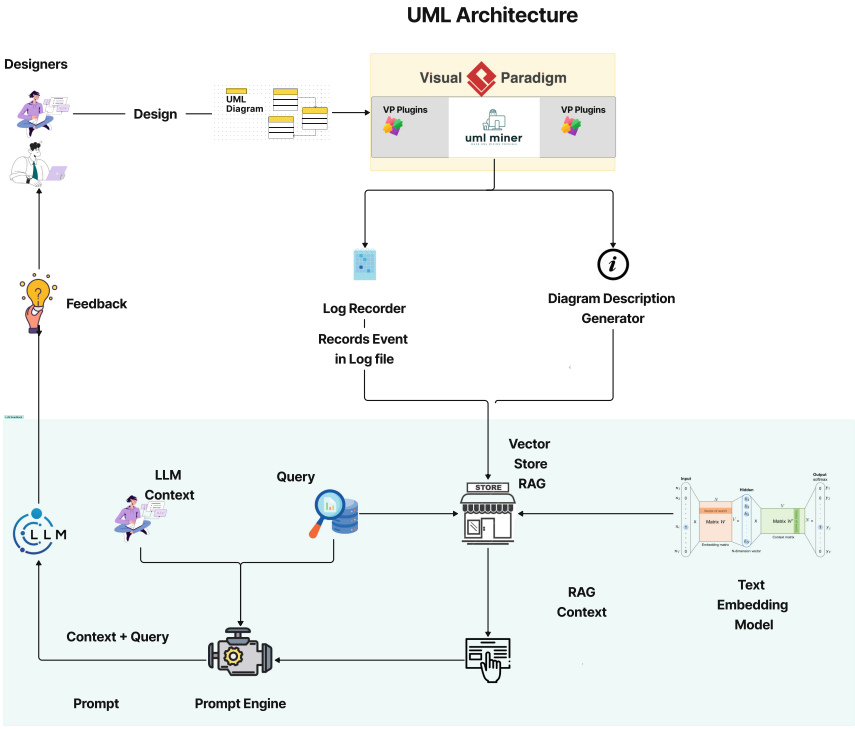


Fig. 1. Architecture of the UML Miner Plugin

feedback that enhances learning outcomes. Our solution was integrated into the existing UML Miner plugin [3] for Visual Paradigm, introducing key extensions:

- **Contextual understanding of design choices:** Generates real-time natural language descriptions of diagrams, enabling RAG-LLM to assess alignment with conceptual and functional goals beyond structural matching.
- **Clarification of abstract concepts:** Enhances feedback by better interpreting abstract UML elements, improving students' understanding of modeling principles.

The remainder of this paper is structured as follows: Sect. 2 introduces the tool, while Sect. 3 presents a running example based on a student-generated solution from the study. Finally, Sect. 4 summarizes the findings and discusses directions for future research.

2 UML Miner

The architecture of the UML Miner plugin is shown in Fig. 1. All components, except the LLM Feedback module, are integrated within the Visual Paradigm

environment. The LLM Feedback module runs as a standalone server and leverages LLaMA 3.1 8b [6] as its underlying language model. The following subsections describe each component in detail.

2.1 Log Recorder

The Log Recorder captures only UML modeling events, creating a unique log for each "Visual Paradigm" project across all sessions. A session includes events from project opening to closure. If a project is reopened, UML Miner appends new activities to the existing log, ensuring a complete historical record. Modeling events are stored in the IEEE-standard XES format. They include actions affecting the project (e.g., adding/removing diagrams, opening, saving), UML diagrams (e.g., adding/removing elements), model properties (e.g., visibility, attribute names), relationships (e.g., associations, aggregations), and extension mechanisms (e.g., stereotypes, tagged values). Events related to style and formatting are excluded. To enable integration with process mining tools, a dedicated language was developed for event representation [2]. Additional details may be recorded depending on the UML element involved.

2.2 Diagram Description Generator

This component converts UML class diagrams into human-readable descriptions. When a student requests feedback, their diagram is transformed and sent to the LLM component. This step is essential to optimize the RAG-LLM's feedback for several reasons:

- **Contextual understanding of design choices**: The natural language description explicitly details the elements and relationships in the diagram, ensuring that the RAG-LLM can assess whether the student's solution aligns with the assignment requirements beyond mere structural comparisons.
- **Clarification of abstract concepts**: UML diagrams often encapsulate abstract concepts that may not be immediately evident. A structured textual representation, such as the one derived from the provided XML schema, makes these elements more explicit, enabling the LLM to better interpret their intent and function.
- **Feedback generation**: By leveraging a well-structured representation, the RAG-LLM can identify missing relationships, incorrect attributes, and inconsistencies, leading to tailored feedback that enhances diagram clarity and adherence to best design practices.
- **Iterative learning process**: By comparing the student's diagram with a knowledge base stored in the vector store, the RAG-LLM can highlight learning opportunities, enabling students to refine their diagrams based on constructive, data-driven insights.

2.3 LLM Feedback

The RAG-LLM employs a specialized model tailored for the UML domain to generate text embeddings that represent UML diagram information. These embeddings are stored in the RAG vector repository, which serves as a centralized knowledge base containing numerical representations extracted by the Diagram Description Generator. The LLM Feedback component, depicted in Fig. 1, consists of the following key elements:

- **Text Embedding Model**. This model converts UML diagram data into numerical embeddings, capturing semantic relationships and structural details. These embeddings facilitate meaningful analysis and enable comparisons between different UML diagrams.
- **Vector Store RAG**. A repository that stores and organizes embeddings produced by the Text Embedding Model. It enables efficient semantic retrieval of UML-related data and supports both corrective and educational feedback to enhance students' understanding of class diagram modeling.
- **Context**. The system retrieves contextual information from both the UML Miner plugin and the Vector Store RAG, ensuring that feedback aligns with the student's query. The context includes the natural language description of the student's solution, enhancing the relevance and accuracy of the responses.
- **Query**. A structured input that includes the user's query about their UML diagram, along with relevant identifiers for the project, diagram, and conversation. The query serves as the main input guiding the retrieval of pertinent knowledge and response generation.
- **Prompt Engine**. This component calculates the cosine similarity between the user's query and stored embeddings, identifying the most relevant information. Additionally, it constructs the final prompt by combining retrieved context, user input, and the LLM's knowledge base.
- **Prompt: Context + Query**. The final structured input provided to the LLM, integrating the retrieved context with the user's specific inquiry. This ensures coherent and contextually relevant responses tailored to the student's needs.
- **LLM**. The core language model, enhanced with UML-specific knowledge through a one-time fine-tuning process. The fine-tuning incorporated sources such as UML 2.5 specifications and academic resources like *The Elements of UML 2.0 Style* from Cambridge University Press. This enables the model to generate precise, domain-specific responses for UML diagram-related queries.

The workflow for executing a question-answering task in this system follows these sequential steps:

1. **Vector Store Initialization**: The system generates embedding vectors for all documents in the knowledge base, indexing them within the Vector Store RAG.
2. **Context Retrieval**: The user's query is transformed into an embedding using the same method applied to stored documents. The system retrieves

the most relevant context from the Vector Store to provide a solid basis for generating a response.
3. **Response Generation**: The retrieved context, along with the user's query, is fed into the LLM. This ensures that responses are not only accurate but also aligned with UML-specific information contained in the local knowledge base.

3 Running Example

The feedback component supports both novices and advanced users by offering predefined and open-ended queries throughout the modeling process. To ensure alignment with learning objectives, instructors define the task in natural language, which guides both students and the RAG-LLM in generating context-aware, iterative feedback based on the specified requirements. Figure 2 shows a student-generated diagram that does not fully meet the assignment requirements. In response to the question "Provide suggestions for adding processes or contents to this diagram", the LLM produced detailed recommendations to deepen the student's understanding of UML modeling concepts.

1. **Understanding entity relationships:** The feedback suggests introducing a `DegreeCourse` class to represent degree programs. By doing so, the student learns how to model entities that structure academic domains and how to define associations between students and their enrolled courses. The feedback also introduces relevant attributes (`courseName`) and operations (`isEnrollmentOpen()`, `closeEnrollment()`), reinforcing the concept of encapsulating behaviors within domain-specific entities.
2. **Clarifying Generalization and Specialization:** The student is suggested to add a `Professor` class related to `DegreeCourse` via an appropriate relationship. This helps illustrate the concept of generalization versus association in UML, guiding the student in choosing the correct modeling approach based on real-world semantics.
3. **Defining Object responsibilities:** The feedback introduces the method `hasSubmittedExam(exam: Exam): boolean` in the `Student` class, helping the student understand how to properly assign responsibilities to objects and implement behavior-driven modeling. This recommendation clarifies the role of operations in encapsulating state-related logic within domain objects.
4. **Tracking Object state and lifecycle:** The recommendation to introduce an attribute `receivedCorrection` in the `Student` class or to create an auxiliary object for tracking exam submissions teaches the student how to manage object state over time. This guidance strengthens the understanding of persistence and state transitions within an object-oriented model.
5. **Refining behavioral constraints:** The feedback suggests restructuring the completeness check by shifting the `isActivityComplete()` method from `Activity` to `Student`. This correction emphasizes the importance of placing behavioral constraints in the most relevant class, helping the student refine responsibility allocation within the model.

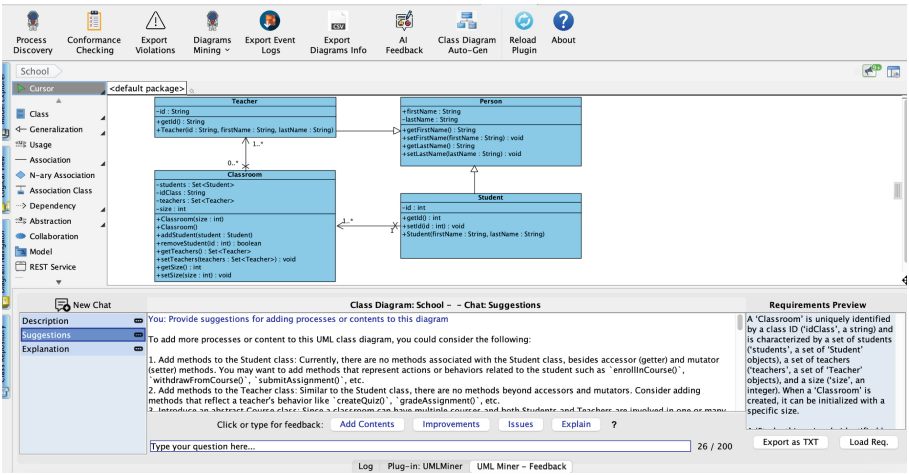


Fig. 2. Feedback example for the predefined question: 'Suggestions for adding processes or contents to this diagram.'

6. **Improving model organization with Abstraction:** The introduction of an abstract superclass `HumanPerson`, encapsulating shared attributes of `Person`, `Student`, and `Professor`, helps the student recognize the advantages of abstraction. This guidance supports the development of a more structured and maintainable class hierarchy.

These recommendations not only improve model completeness but also deepen students' understanding of object-oriented modeling by reinforcing key UML principles and supporting the creation of well-structured, meaningful diagrams.

4 Conclusions

This study extends our ongoing work on an AI-driven feedback system that supports UML class diagram learning. Leveraging a RAG-LLM, the tool delivers personalized, context-aware guidance to help students iteratively improve their models. Future work includes a thorough experimental evaluation of feedback effectiveness and the extension of the tool to additional UML diagram types to broaden its applicability. We also aim to investigate whether students can transfer acquired knowledge to new scenarios without AI support.

Although helpful, AI-generated feedback can lead to over-reliance if not carefully managed. Future studies and improvements to the tool will focus on encouraging students to think independently and apply their knowledge without relying on AI. The RAG-LLM approach also presents limitations. Its performance depends on a well-structured, domain-specific knowledge base, which is costly

to build and maintain. Additionally, better adaptation to diverse learner profiles is needed. These challenges point to future improvements in scalability and pedagogical value.

References

1. Alhazmi, S., Thevathayan, C., Hamilton, M.: Learning UML sequence diagrams with a new constructivist pedagogical tool: SD4ED. In: Sherriff, M., Merkle, L.D., Cutter, P.A., Monge, A.E., Sheard, J. (eds.) SIGCSE 2021: The 52nd ACM Technical Symposium on Computer Science Education, Virtual Event, USA, March 13–20, 2021, pp. 893–899. ACM (2021). https://doi.org/10.1145/3408877.3432521
2. Ardimento, P., Aversano, L., Bernardi, M.L., Carella, V.A., Cimitile, M., Scalera, M.: UML miner: a tool for mining UML diagrams. In: ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, MODELS 2023 Companion, Västerås, Sweden, October 1–6, 2023, pp. 30–34. IEEE, Västerås, Sweden (2023).https://doi.org/10.1109/MODELS-C59198.2023.00014
3. Ardimento, P., Bernardi, M.L., Cimitile, M., Scalera, M.: A rag-based feedback tool to augment UML class diagram learning. In: MODELS Companion 2024, Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems, pp. 26–30. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3652620.3687784
4. Crahen, E., Alphonce, C., Ventura, P.: QuickUML: a beginner's UML tool. In: OOPSLA 2002, Companion of the 17th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, pp. 62–63. Association for Computing Machinery, New York (2002). https://doi.org/10.1145/985072.985105
5. Foss, S., Urazova, T., Lawrence, R.: Learning UML database design and modeling with autoer. In: Kühn, T., Sousa, V. (eds.) Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, MODELS 2022, Montreal, Quebec, Canada, October 23–28, 2022, pp. 42–45. ACM (2022). https://doi.org/10.1145/3550356.3559091
6. Touvron, H., et al.: Llama: open and efficient foundation language models (2023)

A Zero-Shot LLM Framework for Automated Assignment Grading in Introductory Higher Education Courses

Calvin Yeung[1,3(✉)], Jeff Yu[1], King Chau Cheung[1], Tat Wing Wong[1], Chun Man Chan[1], Kin Chi Wong[2], and Keisuke Fujii[3]

[1] Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China
{kingchaucheung,tw.wong,chunmanchan}@cuhk.edu.hk
[2] Centre for Learning Enhancement And Research, The Chinese University of Hong Kong, Hong Kong, China
kc.wong@cuhk.edu.hk
[3] Graduate School of Informatics, Nagoya University, Nagoya, Japan
{yeung.chikwong,fujii}@g.sp.m.is.nagoya-u.ac.jp

Abstract. Automated grading has become a crucial tool in education technology, efficiently assessing large volumes of student work, offering consistent evaluations, and providing immediate feedback. However, current systems face challenges such as the need for large datasets, lack of personalized feedback, and a focus on benchmark performance over student experience. To address these issues, we propose a Zero-Shot Large Language Model (LLM)-Based Automated Assignment Grading (AAG) system. This framework uses prompt engineering to evaluate both computational and explanatory responses without additional training or fine-tuning. The system delivers tailored feedback, highlighting strengths and areas for improvement, thus enhancing learning outcomes. Our study shows the system's effectiveness through evaluations, including survey responses from introductory higher education students, which indicate significant improvements in motivation, understanding, and preparedness compared to traditional grading methods. These results demonstrate the AAG system's potential to transform educational assessment by focusing on learning experiences and providing scalable, high-quality feedback.

Keywords: Automated Grading · Zero-Shot Learning · Large Language Models · Personalized Feedback · Education Technology

1 Introduction

In the field of education technology, automated grading has long been a key objective due to its ability to efficiently assess large volumes of work [4], deliver consistent and unbiased evaluations [8], and provide immediate feedback to enhance student learning [11]. It also allows educators to focus on developing engaging learning experiences and leverage data-driven insights for improved teaching strategies.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 19–27, 2025.
https://doi.org/10.1007/978-3-031-99264-3_3

Automated grading tools were first introduced in the 1960 s for programming [12] and essay evaluation [18]. Since then, various traditional methods–such as unit testing, rule-based approaches, and domain-adaptive models–have been proposed to compare student submissions against reference answers using text similarity, predefined rules, and measurable features [10,15,16]. With the advent of deep learning, neural models such as LSTMs [17,19], CNNs [20], and prototypical neural networks [23] have shown improved performance over traditional approaches. These systems, however, often prioritize accuracy over interpretability and student-centered outcomes.

The rapid advancement of Large Language Models (LLMs) in natural language processing has opened new directions for automated grading. LLMs such as BERT [5], GPT-3 [3], and GPT-4 [1] have demonstrated strong capabilities in question answering and summarization tasks without requiring task-specific training. Consequently, few-shot learning approaches–where limited labeled examples guide the model–have become a popular method in LLM-based grading systems. Despite their success, these methods still require curated datasets for each new domain or course, which can be time-consuming and impractical for dynamic educational content.

Zero-shot learning, by contrast, enables LLMs to generalize to new tasks using only task instructions in the form of prompts. While preliminary work has explored the feasibility of zero-shot LLMs in grading [14], results have been mixed, and simple prompting techniques have not yielded robust performance across diverse assignment types. Moreover, automated grading systems often lack mechanisms for delivering personalized feedback, which is crucial for helping students identify specific mistakes and improve. Student perceptions of such systems also remain underexplored; concerns include trust in automated evaluations, lack of transparency, and challenges in interpreting feedback [2,7,13].

To address the limitations of traditional grading, we propose (a) a **Zero-Shot LLM-Based Automated Assignment Grading (AAG) system**, illustrated in Fig. 1, which leverages prompt engineering to assess both computational and explanatory responses without requiring additional training or fine-tuning; (b) a **Tailored Feedback mechanism** that provides personalized, actionable suggestions to enhance student understanding and performance; and (c) an **evaluation based on Student-Centered Metrics**, demonstrating improved engagement and preparedness through survey results from undergraduate and graduate students.

2 Zero-Shot LLM-Based Automated Assignment Grading System

This section introduces the Zero-Shot LLM-Based Automated Assignment Grading (AAG) system. An overview of the system is illustrated in Fig. 1. By automatically providing feedback to both teachers and students, the AAG system could significantly reduce the manual grading workload while offering comprehensive and detailed insights to support continuous improvement. The AAG system comprises the following core components:

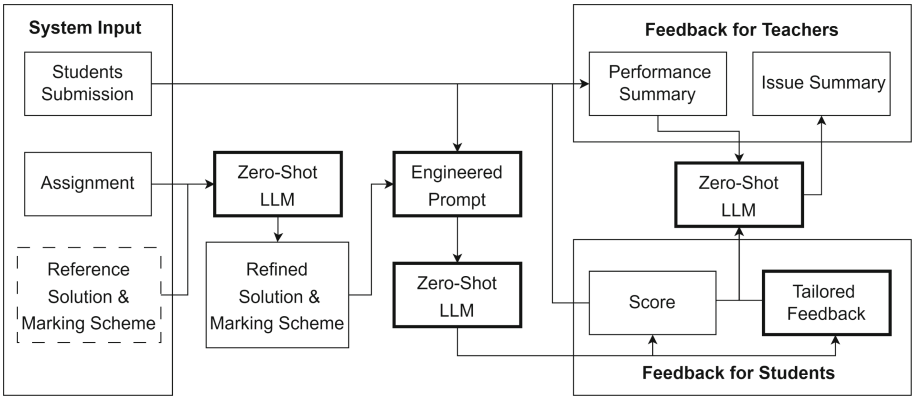


Fig. 1. Overview of the Zero-Shot LLM-Based AAG System. The bolded box highlights the core innovation of this study: an AAG system integrating prompt engineering, zero-shot LLM capabilities, and tailored feedback for students. The dashed line box indicates non-complementary input.

Adaptive Input Processing and Marking Scheme Refinement: The AAG system processes three key inputs: the student's assignment submission, the assignment questions, and optionally (indicated by the dashed box in Fig. 1), a reference solution or marking scheme. While the reference solution and marking scheme can provide additional guidance, they are not essential for grading.

The system also generates or refines scoring guidelines based on the assignment questions, which is particularly useful for introductory courses with assignments tied to timely topics (see Table 1). This flexibility enables instructors to create diverse, engaging questions without the need for constantly updating marking schemes.

Leveraging LLMs for Automated Grading: The AAG system uses GPT-4 [1] for solution refinement, student submission evaluation, and summarizing student issues. Among the leading LLMs–such as Llama-3 [6], Qwen-2 [22], Claude-3, and Gemini-1.5 [21]–GPT-4 was found to be the most effective based on a comparative evaluation using identical prompts.

The system integrates the student submission, assignment questions, and marking scheme into the evaluation prompt, as shown in Fig. 2. The prompt outlines the task, specifies the course context, provides the assignment details, and highlights the relevant question or subquestion. It also includes the scoring scale (1 to 10) and feedback type needed. With a well-engineered prompt and refined marking scheme, the LLM can generate effective feedback without the need for fine-tuning.

Prompt for LLM

Overview: Your role is to assess and provide feedback on a student's response to a specific task in a <Course> assignment. Each task contains multiple components, and you are required to focus on evaluating the last question.
Background: <Question Background>
Question: <Question>
Student's Response: <Student Answer>
Marking Scheme: <Marking Scheme>

Instructions for the Model:

- **Focus:** Grade and provide comments exclusively on the last question.
- **Scoring:** Rate the student's answer from 1 to 10 according to the provided marking scheme.
- **Feedback:** Provide a brief and constructive critique. Highlighting strengths and areas for improvement according to the marking scheme. Feedback must be concise.

Fig. 2. Evaluation prompt for the AAG system. Angle brackets < > mark placeholders. A single question may include subquestions and reference prior answers. The prompt captures full context but focuses on evaluating the latest question and answer.

Personalized Feedback for Students: Students receive personalized feedback with a score and detailed comments highlighting errors, explaining misconceptions, and offering improvement suggestions. Unlike traditional grading, which often includes only brief remarks, this feedback provides actionable insights to deepen understanding and their approach to questions answering.

Performance Overview for Teachers: For teachers, the system generates a "Performance Summary" that offers an overview of class and individual performance relative to assignment objectives. It includes summary statistics, question-level breakdowns, and common student issues identified through LLM-generated feedback, providing actionable insights to inform instruction and support student progress.

Limitations of the AAG System: Despite the strengths of the AAG system. Grading consistency depends on prompt design and marking quality, though human-guided automation may alleviate this. The absence of iterative feedback may limit learning, but multi-stage feedback could address this gap.

3 Experiments and Results

This section first evaluates the AAG system's effectiveness by comparing its grading with human TAs. Afterward, students' feedback is included to assess

perceived usefulness. The experiments adheres to university ethical guidelines, ensuring privacy, confidentiality, and informed consent. More information available at https://github.com/calvinyeungck/Automated_Assignment_Grading.

3.1 AAG Grading Evaluation

While substantial evidence [1] supports the effectiveness of LLMs in grading, a limitation of LLM-based grading lies in the distributional differences between AI scorers and humans [9]. To address this, we compared scores for 150 student assignments graded by TAs and the AAG system on two open-ended questions from the "Introduction to Statistics" course. Open-ended questions were chosen for their flexibility in student responses, unlike calculation-based questions at the introductory level. A summary of these questions is provided in Tables 1.

Table 1. Background and Question for "Introduction to Statistics" Assignment.

Background	You are a Statistician at the Census and Statistics Department, tasked with leading a team to collect data on the aging population in a housing estate in <anoynomised location>. The collected data will be reported to the Social Welfare Department.
Question 1	What kind of sampling method will you suggest using?
Background	You would like to conduct a survey to predict the results of the US presidential election.
Question 2	Find a good variable for forming strata. Explain your choice briefly.

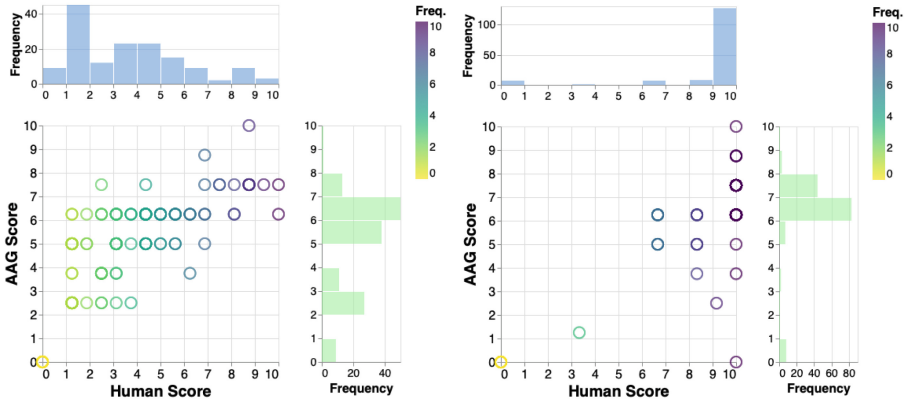


Fig. 3. Human and AAG system grading distribution of Q1 (left) and Q2 (right).

Figure 3 shows the grading distributions for Question 1 (left) and Question 2 (right) by human graders and the AAG system. Pearson correlation coefficients between human and AAG scores for Questions 1 and 2 were 0.75 and 0.82, respectively, indicating strong correlations ($p < 0.0001$) and suggesting the AAG system closely aligns with human grading.

Despite high correlations, score distributions differed between questions. While both were open-ended, the marking schemes varied significantly: Q1 had detailed guidelines, while Q2 only had general criteria. As a result, scores for Q1 were more diverse, and those for Q2 were more concentrated. This highlights the role of a refined marking scheme in ensuring consistent and accurate grading.

A qualitative analysis of scoring discrepancies revealed that some differences were due to human factors, such as grading errors or failure to follow marking guidelines. These inconsistencies emphasize the need for the AAG system to improve grading consistency and provide better feedback.

3.2 Student Perspective on AAG System

Survey Design: To evaluate the AAG system's effectiveness, a voluntary survey was conducted with students who received feedback from both the AAG system and TAs. Participants included students from the undergraduate course "Introduction to Statistics" and the graduate courses "Data Management" and "Principles of Risk Management" at The Chinese University of Hong Kong. They were informed that their responses would not affect their assignment scores. The survey consisted of 10 questions assessing the quality of the feedback and its impact on learning. Questions 1 to 9 used a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), and Question 10 asked students to choose between traditional TA grading and AAG feedback.

Survey Questions: The survey questions targeted multiple dimensions of the feedback experience. **Q1-Q2:** Focused on the clarity and helpfulness of the feedback in identifying specific mistakes. **Q3-Q4:** Evaluated how well the feedback clarified why answers were incorrect and the degree of customization to individual responses. **Q5-Q6:** Measured how effectively the feedback provided actionable steps for improvement and addressed personal learning needs. **Q7-Q8:** Assessed the impact of feedback on future problem-solving and its insight into knowledge gaps. **Q9:** Investigated whether the feedback encouraged motivation to learn and correct mistakes. **Q10:** Asked students to choose between conventional TA grading and the AAG system feedback.

Results and Interpretation: A total of 104 responses were collected, with a Wilcoxon Signed-Rank Test and a Binomial Test applied to Questions 1-9 ($\mu = 3.6 \pm 0.1$) and Question 10 (89% preferred AAG system), respectively. Both tests showed strong statistical significance ($p < 0.0001$), revealing consistently positive feedback. Students reported that the AAG system improved their understanding of mistakes, provided clearer explanations, and offered highly personalized, actionable feedback that motivated learning. They felt better prepared

for future problems due to the system's insights into their learning gaps. The majority of student preferred the AAG system over traditional TA grading for its individualized feedback.

Additionally, responses from weaker (bottom 50% scorer) and stronger (top 50% scorer) student groups were compared. Weaker students rated the AAG system more positively across all aspects, with a Mann-Whitney U test showing significant differences ($p < 0.05$). Weaker students particularly appreciated the system's help in identifying mistakes, offering improvement strategies, preparing for future problems, and providing encouragement for continuous learning.

4 Discussion

While our AAG framework performs well on structured assignments, its application to ill-defined tasks (e.g., essays or creative problem-solving) could present challenges such as subjectivity and diverse solutions. Nonetheless, the system's modular design allows for potential adaptation through flexible prompts and human-in-the-loop mechanisms. Additionally, traditional assessment methods like rubric-based or TA grading could benefit from integrating AAG components–such as error detection and feedback generation, potentially improving grading consistency, reducing workload, and enhancing fairness and scalability in educational assessment.

5 Conclusion

This paper presents an AAG system with a zero-shot LLM framework to evaluate assignments effectively without additional training or datasets. The system provides personalized feedback, helping students identify strengths and areas for improvement, thereby enhancing learning outcomes. Compared to human grading, the AAG system offers more consistent and higher-quality feedback. Survey results indicate that the tailored feedback from the AAG system significantly boosts motivation, understanding, and preparedness, outperforming traditional grading methods.

Acknowledgments. Funded by the HKUGC FITE, 2023-2026.

Disclosure of Interests. The authors have no competing interests.

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Azad, S., Chen, B., Fowler, M., West, M., Zilles, C.: Strategies for deploying unreliable AI graders in high-transparency high-stakes exams. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 16–28. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_2

3. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
4. Burrows, S., D'Souza, D.: Management of teaching in a complex setting. In: Proceedings of the Second Melbourne Computing Education Conventicle, pp. 1–8 (2005)
5. Chen, S., Lan, Y., Yuan, Z.: A multi-task automated assessment system for essay scoring. In: International Conference on Artificial Intelligence in Education, pp. 276–283. Springer (2024)
6. Dubey, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
7. Grimes, D., Warschauer, M.: Utility in a fallible tool: a multi-site case study of automated writing evaluation. J. Technol. Learn. Assess. **8**(6) (2010)
8. Haley, D., Thomas, P., De Roeck, A., Petre, M.: Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML (2007)
9. Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., Kedzie, C.: LLM-rubric: a multidimensional, calibrated approach to automated evaluation of natural language texts. arXiv preprint arXiv:2501.00274 (2024)
10. Heilman, M., Madnani, N.: ETS: domain adaptation and stacking for short answer scoring. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 275–279 (2013)
11. Hirschman, L.: Automated grading of short-answer tests. IEEE Intell. Syst. Trends Controversies Sect. **15**(5), 22–37 (2000)
12. Hollingsworth, J.: Automatic graders for programming classes. Commun. ACM **3**(10), 528–529 (1960)
13. Hsu, S., Li, T.W., Zhang, Z., Fowler, M., Zilles, C., Karahalios, K.: Attitudes surrounding an imperfect AI autograder. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2021)
14. Kortemeyer, G.: Can an AI-tool grade assignments in an introductory physics course? arXiv preprint arXiv:2304.11221 (2023)
15. Liu, X., Wang, S., Wang, P., Wu, D.: Automatic grading of programming assignments: an approach based on formal semantics. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET), pp. 126–137. IEEE (2019)
16. Messer, M., Brown, N.C., Kölling, M., Shi, M.: Automated grading and feedback tools for programming education: a systematic review. ACM Trans. Comput. Educ. **24**(1), 1–43 (2024)
17. Mizumoto, T., et al.: Analytic score prediction and justification identification in automated short answer scoring. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 316–325 (2019)
18. Page, E.B.: The imminence of... grading essays by computer. The Phi Delta Kappan **47**(5), 238–243 (1966)
19. Riordan, B., Horbach, A., Cahill, A., Zesch, T., Lee, C.: Investigating neural architectures for short answer scoring. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 159–168 (2017)
20. Shin, J., Gierl, M.J.: More efficient processes for creating automated essay scoring frameworks: a demonstration of two algorithms. Lang. Test. **38**(2), 247–272 (2021)

21. Team, G., et al.: Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
22. Yang, A., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
23. Zeng, Z., Li, L., Guan, Q., Gašević, D., Chen, G.: Generalizable automatic short answer scoring via prototypical neural network. In: International Conference on Artificial Intelligence in Education, pp. 438–449. Springer (2023)



From Code to Insight: How LLMs Help and Hinder Qualitative Research

Chaewon Kim[1(✉)], Fengfeng Ke[2(✉)], Alex Barrett[1(✉)], and Nuodi Zhang[1(✉)]

[1] Florida State University, Tallahassee, FL 32311, USA
{ck22j,abarrett3,nzhang4}@fsu.edu
[2] University of Maryland, College Park, MD 20742, USA
fke@umd.edu

Abstract. This study evaluates the performance of large language models (LLMs) in the qualitative analysis of an interview transcript. Four state-of-the-art LLMs—GPT-4, LLAMA 3.1, Gemini 2.0 flash, and DeepSeek-V3—were compared with human qualitative analysis using both inductive and deductive approaches. The deductive analysis employed the socio-ecological model as a framework to examine the complexities of supporting neurodivergent individuals in rural communities. Results indicate that Gemini achieved the highest cosine similarity to human coding in the inductive approach, while DeepSeek performed best in the deductive approach. Graph neural network (GNN) visualizations revealed that certain models struggled to capture the holistic context of the interview, demonstrating limitations in comprehension and contextual analysis compared to human coders. The models used in this study were freely available, and participant privacy was protected by anonymizing the transcript. The study was approved by the institutional review board. These findings highlight both the potential and the challenges of employing LLMs to augment qualitative research, particularly in nuanced and context-dependent data analyses. We discuss the implications of using LLMs to enhance qualitative research processes and propose future directions for improving their alignment with human analytical processes.

Keywords: Qualitative Research · Large Language Models (LLM) · Human-AI Collaboration

1 Introduction

[8] argue that mental health practitioners, including psychiatrists, must be well-versed in multiple paradigms that shape the understanding of human experiences. Relying solely on one perspective limits both the depth and breadth of insight. This principle extends to research, where embracing diverse methodologies enhances our comprehension of complex phenomena. Qualitative research plays a crucial role in capturing the richness and complexity of human experiences, providing context-sensitive insights that quantitative approaches alone may overlook.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 28–35, 2025.
https://doi.org/10.1007/978-3-031-99264-3_4

The quality of qualitative research is judged by its validity, reliability, and transferability [11]. Credibility, akin to internal validity in quantitative research, ensures that findings are trustworthy from the participants' perspectives. Transferability, corresponding to external validity, speaks to the applicability of findings beyond the immediate study context. Dependability and confirmability, which can be operationalized as reliability in qualitative research [11], further strengthen the trustworthiness of qualitative inquiry.

However, achieving and demonstrating this level of rigor is no small feat—it demands reflexivity, critical thinking, and a philosophical commitment to naturalistic inquiry. Given this inherent challenges of conducting high-quality qualitative research, there is growing interest in leveraging AI, particularly Large Language Models (LLMs), to complement human researchers in qualitative analysis. LLMs excel at processing vast amounts of text, identifying patterns, and generating coherent narratives [6][14]. Their ability to handle complex linguistic structures has positioned them as valuable tools in fields such as natural language processing and content generation.

Yet, the question remains: Can LLMs truly contribute to the intellectual rigor required for qualitative research? Unlike simple tasks such as summarization or paraphrasing, qualitative inquiry demands deep contextual understanding, sensitivity to cultural and social nuances, and the ability to interpret complex human experiences [7]. Moreover, qualitative analysis often requires a level of creativity, intuition, and theoretical engagement that AI systems may struggle to replicate [4]. The integration of LLMs into qualitative research is an area of active exploration. While AI can assist with certain procedural aspects—such as data organization, initial coding, and pattern recognition—the essence of qualitative inquiry lies in meaning-making, interpretation, and critical engagement, all of which remain deeply human endeavors.

This proposal contributes to this ongoing discourse by examining the potential and limitations of AI collaboration in qualitative research, seeking to understand how these technologies might assist—not substitute—the intellectual and methodological rigor of human, which is fundamental to the field.

2 Related Work

2.1 The Ecological Systems Model

An underlying theoretical lens for the deductive coding in our qualitative inquiry is the ecological systems theory [5]. The theory was widely used in framing various research in education, social work, and healthcare to provide a holistic understanding of different systems influencing human experience [3, 12]. The theory consists of five interrelated systems: microsystem, mesosystem, exosystem, macrosystem, and chronosystem. The microsystem includes the individual actors and immediate environments of individuals, such as teachers, peers, classroom culture, and resources, shaping both formal and informal learning. The mesosystem highlights the dynamic relationships between these micro-level factors, emphasizing their interconnected and evolving nature. The exosystem comprises external influences like school leadership, policies, and community involvement, indirectly affecting the learner. The macrosystem extends further to societal, political, and educational contexts, including standardization, accountability, and curriculum mandates that shape the broader educational environment. Finally, the

chronosystem accounts for the influence of time, considering how a learner's schooling experience evolves throughout their primary and secondary education.

2.2 Using LLM (NLP) for Qualitative Coding

Natural Language Processing (NLP) has become a valuable tool in qualitative coding in recent years, allowing researchers to process large volumes of unstructured text data with greater efficiency. Traditional qualitative coding is labor-intensive and constrained by the amount of data that can be manually analyzed. NLP techniques and off-the-shelf qualitative software with NLP functions can significantly reduce research time and cost [1]. Large Language Models (LLMs), like GPT-3.5, further enhance qualitative analysis by generating detailed codes and identifying broad patterns across extensive datasets [9]. However, while NLP accelerates data processing and expands analytical scope, it does not fully replace the depth of human interpretation [1]. Subtleties such as contextual nuance in qualitative research remain challenging for automated systems [9]. As a result, NLP serves as a complement rather than a substitute, streamlining qualitative analysis while leaving space for human expertise in interpreting complex social phenomena.

3 Method

3.1 Context

As part of a bigger project on analyzing the need to design a care ecosystem for neurodivergent populations in rural areas in the U.S., the researchers interviewed 28 teachers, caregivers, and healthcare professionals. For this study, we chose an interview transcript from a teacher who is also a caregiver of a child with ADHD residing in a rural area. The semi-structured interview lasted for about an hour, with core questions on how they navigate the challenges in supporting neurodivergent students. Due to the complex insights from both roles as a teacher and a caregiver, we started exploring how to leverage LLM to deepen our understanding of the given transcript. This project was approved by the institutional review board and the transcript was anonymized to protect the participant's privacy.

The researchers first coded the transcript using both inductive and deductive approaches for thematic analysis [13]. Inductive coding means that the themes emerged from interpreting the transcript without a guiding framework, while the ecological systems model served as a lens to interpret the transcript in deductive manner [5]. The same prompt was used to obtain the qualitative analysis from four state-of-the-art generative LLMs: GPT-4, LLAMA 3.1, Gemini 2.0 flash, and DeepSeek-V3. All models were free to use, and the transcript was analyzed with the personalization setting turned off.

3.2 Data Analysis

Through inductive and deductive coding via human and four state-of-the-art LLMs, we obtained 10 text files from the qualitative analysis of the selected interview transcript. The code analyzes textual outputs from all five sources by loading text files,

converting sentences into 386 dimensional embeddings using the SentenceTransformer (all-MiniLM-L6-v2) model. The cosine similarity was calculated between the TF-IDF vector representations of the distinct text files, measuring the similarity of their overall content. A Python package called NetworkX was used to construct a similarity-based graph (Fig. 1). TF-IDF was applied to extract key terms for each source and reduce the embedding dimensions using UMAP (with an option for t-SNE) for visualization.

4 Results

4.1 Cosine Similarities and Visualizations of Inductive and Deductive Approaches

The threshold of interpreting the meaningfulness of cosine similarities is relative, pending on the context [10]. Overall, Gemini showed the highest similarity with the human analysis in the inductive approach, while DeepSeek outperformed the other LLM models in the deductive approach. The similarity-based graphs in both approaches show that some LLM models do not fully capture the complexity of the given transcript as the human and other LLM models do. Specifically, it is observed that the output from LLAMA was only distributed around below 5 on dimension 2, unlike other analyses (see Fig. 1).

Table 1. Cosine similarities of LLM-generated and human qualitative analysis

Approach	Models					
		GPT	LLAMA	Gemini	DeepSeek	Human
Inductive	GPT	1				
	LLAMA	.494	1			
	Gemini	.696	.498	1		
	DeepSeek	.492	.339	.537	1	
	Human	.413	.343	.510	.441	1
Deductive	GPT	1				
	LLAMA	.525	1			
	Gemini	.618	.589	1		
	DeepSeek	.703	.590	.694	1	
	Human	.445	.424	.539	.558	1

4.2 LLM-Generated Outliers

There was a section where human data points were absent, but numerous AI-generated texts were present in GNN visualizations. In Fig. 1, human text did not appear beyond 12 on Dimension 1, whereas ChatGPT (n = 13), LLAMA (n = 1), Gemini (n = 9),

and DeepSeek (n = 16) all exceeded this threshold. In Fig. 2, a distinct cluster formed by ChatGPT (n = 16), LLAMA (n = 7), Gemini (n = 8), and DeepSeek (n = 15) emerged below 13 on Dimension 1, containing no human-generated data points. This could suggest either that the LLMs are capturing something humans overlook or that they struggle to provide a concise understanding of the given transcript.



Fig. 1. Semantic similarity graph of LLM-generated and human inductive analysis (top) and deductive analysis (bottom)

Examining further the outliers showed that neither was the case. In the inductive analysis, the outliers from LLMs could be categorized as (a) the explanation of the coding procedure (n = 10), (b) titles and subheadings (n = 23), or (c) conversational text asking the user satisfaction (n = 6), none of which the human analysis necessarily

contained. Similarly, in the deductive analysis, only 6 out of 39 outliers contained meaningful information about the themes, while the other 33 being either the explanation of the coding procedure (n = 8) or titles and subheadings (n = 25). The six themes counted as outliers were *behavioral issues affecting classroom dynamics* and *strategies for managing classroom engagement* (from ChatGPT-4), *teacher preparedness and collaboration*, *varying preparedness*, *parent perspective* and *collaborative strategies* (from Gemini 2.0 flash).

The outlier themes, while containing overlapping words from the transcripts and human-coded analysis, lack contextual depth, which can lead to misunderstandings. For example, "*engagement*" was mentioned in the specific context of motivating a student with ADHD to set academic goals, rather than as a general classroom strategy. A more accurate coding than "strategies for managing classroom engagement" would be "setting clear goals for neurodivergent students' growth." Similarly, teacher preparedness should be understood alongside limited professional development and funding for accommodations, as highlighted earlier in the transcript. Without a holistic consideration of its dialogical contexts, there is a risk of misinterpretation, such as wrongly attributing the challenge to teachers rather than systemic constraints.

5 Discussion

The application of topic modeling and NLP techniques has long been a staple in qualitative research, offering efficient means of processing and analyzing large textual datasets. However, a persistent challenge remains in assessing the quality and validity of these automated techniques when applied to the inherently interpretative nature of qualitative research [2]. Unlike quantitative approaches, qualitative research lacks a universally accepted standard for evaluating the appropriateness of themes and categorizations, leading to ambiguity in determining the "desired" quality of analysis. This study compared different LLMs with human performance in both inductive and deductive qualitative research approaches. By doing so, we shed light on the discrepancies and overlaps between machine-generated and human-constructed themes. Notably, the themes that emerged as outliers from LLM analyses highlight the abstractness and lack of contextual depth in LLM-generated qualitative insights, as previously pointed out [9]. These outliers, rather than capturing nuanced human experiences, often oversimplify complexities or misrepresent the core ideas present in the original transcript. For instance, the interviewee's description of classroom engagement strategies was situated in the complaint about supporting individuals with neurodiversity under constrained resources. The AI-generated analysis, however, treated it as a standalone theme, failing to acknowledge its deeper connection to the teacher's broader concerns about inadequate funding and training. Similarly, the AI-generated theme of *parent perspective* ignored its integral relationship with the teacher's dual role as both an educator and a caregiver, overlooking the intersectionality nature of her experiences.

While this study contributes to an ongoing conversation about the utility of LLMs in qualitative research, future investigations should adopt a more systematic approach to refining AI-assisted methodologies. Additionally, it is important to acknowledge that all tested LLMs were publicly available free versions, which may not reflect their full

analytical potential. Thus, this study does not aim to rank or determine the superiority of any specific model but rather to highlight their capabilities and limitations in supporting qualitative analysis.

6 Conclusion

Qualitative research remains a challenging and resource-intensive endeavor due to its reliance on deep contextual interpretation and nuanced meaning-making. This study explored the ways in which LLMs can both help and hinder in qualitative research, demonstrating their potential to streamline certain analytical tasks while exposing their limitations in capturing the depth and nuances of human experiences. By highlighting both the strengths and shortcomings of LLM-driven qualitative analysis, this research contributes to a broader discussion on the evolving role of AI in qualitative research.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Abram, M.D., Mancini, K.T., Parker, R.D.: Methods to integrate natural language processing into qualitative research. Int J Qual Methods **19**, 1609406920984608 (2020)
2. Adams, T.E., Holman Jones, S., Ellis, C.: Autoethnography: Understanding Qualitative Research. Oxford University Press, Oxford (2015)
3. Anderson, J., Boyle, C., Deppeler, J.: The ecology of inclusive education: Reconceptualising Bronfenbrenner. In: Equality in Education: Fairness and Inclusion, pp. 23–34. SensePublishers, Rotterdam (2014)
4. Braun, V., Clarke, V.: Reflecting on reflexive thematic analysis. Qual. Res. Sport Exercise Health **11**(4), 589–597 (2019). https://doi.org/10.1080/2159676X.2019.1628806
5. Bronfenbrenner, U.: Ecological models of human development. Int. Encycl. Educ. **3**(2), 37–43 (1994)
6. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020). https://arxiv.org/abs/2005.14165
7. Creswell, J.W., Poth, C.N.: Qualitative Inquiry and Research Design: Choosing Among Five Approaches, 4th edn. SAGE Publications, Thousand Oaks (2018)
8. Fossey, E., Harvey, C., McDermott, F., Davidson, L.: Understanding and evaluating qualitative research. Aust. N. Z. J. Psychiatry **36**(6), 717–732 (2002). https://doi.org/10.1046/j.1440-1614.2002.01100.x
9. Liu, X., Zhang, J., Barany, A., Pankiewicz, M., Baker, R.S.: Assessing the potential and limits of large language models in qualitative coding. In: International Conference on Quantitative Ethnography, pp. 89–103. Springer, Cham (2024)
10. Majumder, G., Pakray, P., Avendaño, D. E. P.: Interpretable semantic textual similarity using lexical and cosine similarity. In: Social Transformation–Digital Way: 52nd Annual Convention of the Computer Society of India, CSI 2017, Kolkata, India, 19–21 January 2018, Revised Selected Papers, vol. 52, pp. 717–732. Springer, Singapore (2018)
11. Mays, N., Pope, C.: Assessing quality in qualitative research. BMJ **320**(7226), 50–52 (2000). https://doi.org/10.1136/bmj.320.7226.50

12. Onwuegbuzie, A.J., Collins, K.M., Frels, R.K.: Foreword: using bronfenbrenner's ecological systems theory to frame quantitative, qualitative, and mixed research. Int. J. Multiple Res. Approach. **7**(1), 2–8 (2013)
13. Patton, M.Q.: Enhancing the quality and credibility of qualitative analysis. Health Serv. Res. **34**(5 Pt 2), 1189 (1999)
14. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 1–67 (2020). https://arxiv.org/abs/1910.10683



Grading University Students with LLMs: Performance and Acceptance of a Canvas-Based Automation

Luke Korthals(✉), Hannes Rosenbusch, Raoul Grasman, and Ingmar Visser

University of Amsterdam, 1018 WS Amsterdam, The Netherlands
l.korthals@uva.nl

Abstract. Teachers in higher education spend considerable time grading assignments rather than tutoring students. Large language models (LLMs) could address this by generating human-like grades and feedback for assignments. However, accounts of their practical application are scarce. We wrote Python code to integrate the Canvas learning management system with an API for GDPR-compliant LLM access. We used this AI system to grade and feedback weekly assignments of 58 graduate students (47 study participants) enrolled in an introductory programming course. LLM-grading was fast, cost efficient, and relatively accurate: GPT-4o and human graders agreed perfectly in 80% of 6345 evaluated student answers. Human and LLM-grades were positively correlated ($r = .570 - .866$ for each assignment) but humans awarded higher grades. Disagreements occurred because human graders overlooked student mistakes, and GPT-4o graded ambiguous cases stricter and misgraded some questions. Importantly, we think most LLM-grading mistakes can be reconciled by avoiding certain tasks and improving grading rubrics. Most students (77%) found the LLM-generated feedback helpful, and some (33%) appreciated the rapid grading. However, multiple students (30%) stated that receiving preliminary LLM-grades that were lower than their actual grades made them anxious or upset. We conclude that LLM-based grading and feedback can, and likely should, be used to optimize teachers' resources. Canvas users can adopt our pipeline by simply adding their own API keys, Canvas IDs, and assignments: https://github.com/lukekorthals/canvas-llm-integration.

Keywords: AI grading · AI feedback · large language models · field study · pedagogy · effects on learner · proof-of-concept · learning management system

1 Background

Teachers in higher education often spend a lot of time grading assignments despite the limited educational benefit of grades [5,13,14,22]. Feedback on the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 36–43, 2025.
https://doi.org/10.1007/978-3-031-99264-3_5

other hand is very beneficial for students, but takes even more time and is therefore scarce [10,16,20,21]. Recent studies suggest that large language models (LLMs) can generate accurate grades [7–9,11] and useful feedback [6,12,15]. Leveraging LLM-grading could save teachers time, which may be reinvested directly into students, for example by tutoring those who struggle the most. Additionally, LLM-feedback might improve the experience of students who would otherwise receive no, little, or delayed feedback. However, despite promising pioneer work, actual implementations of LLMs at universities remain scarce. To address this, we created a pipeline for AI grading and feedback and tested it in an introductory programming course in January 2025. Here we share insights from setting up the pipeline and initial findings on grading accuracy and student reception.

2 Pipeline for AI Grading and Feedback

2.1 Overview of the Pipeline

We wrote Python [19] code to integrate *Canvas* [1], a learning management system, with the *OpenAI* API [2]. Importantly, we used an endpoint provided by the *Teaching and Learning Center Science* for internal use at the University of Amsterdam that stores data on GDPR-compliant servers [3,4].

Based on course and assignment IDs the pipeline first downloads students' file submissions from Canvas. Then, each submission is chunked into individual tasks. Next, it prompts an LLM to grade and feedback each task separately. For grading, we included the task, example solution, student answer, and rubrics in the prompt; and instructed the LLM to award zero to one point(s) according to the rubrics (e.g., "subtract 0.25 points if X"). For feedback, the prompt included the task and student answer, and instructed the LLM to give helpful, specific, and detailed feedback. After generating grades and feedback for each task, we instructed the LLM to generate an overall feedback summary, including a custom coding challenge. All prompts instructed the LLM to create a list of step-by-step thoughts and to format its response using html tags (e.g., <feedback>, <points>). This allowed us to retrace the "thoughts" of the model and extract relevant information. In a last step the pipeline creates a markdown file with a complete account of all prompts and LLM responses and a PDF report containing the feedback. Finally, the grades and feedback reports for all students are uploaded to Canvas. All code, prompts, and implementation instructions are available here: https://github.com/lukekorthals/canvas-llm-integration.

2.2 Piloting the Pipeline

In January 2025 we taught R and Python to 58 graduate students enrolled in an introductory programming course at the University of Amsterdam. Spanning four weeks, it included about 30 hours of lectures and tutorials, real-time tutoring via Slack [18], weekly take-home assignments, and one in-class exam.

The course comes with a high workload for students and teachers, and in previous years students waited several days to receive their weekly grades. They also received almost no textual feedback on their submissions, thus providing limited opportunities to improve before the next week's assignment. This year, we hoped to address these concerns by providing students with preliminary grades and feedback generated by GPT-4o[1] [17]. Each student received AI feedback for two randomly chosen assignments. Final grades were awarded by the course instructor and three PhD students who did not see the AI-generated grades before grading.

During the first lecture, we informed students about the course setup and collected written consent from 47 students wo agreed that we could use their data for scientific inquiry. They were between 21 and 42 years old ($M = 24.51$, $sd = 3.78$), most were female ($f = 30$, $m = 14$, *other* $= 1$, $NA = 2$) and most had little previous experience with R (*no experience* $= 12$, *very little* $= 13$, *some* $= 10$, *moderate* $= 9$, *a lot* $= 1$, $NA = 2$) or other programming languages (*no experience* $= 11$, *very little* $= 17$, *some* $= 10$, *moderate* $= 5$, *a lot* $= 2$, $NA = 2$).

3 Findings

3.1 Setting up and Using the Pipeline

Setup. Setting up the integration took considerable time and effort. Without prior knowledge of the APIs, it took the main author about five weeks to write the code and refine rubrics based on test submissions.

Downloading Submissions. We instructed students to structure their file submissions with certain indicators (e.g., "#R1", "#Python2"), such that we could automatically chunk them into question-wise answers. Almost all submissions were correctly formatted, and we only had to manually re-format a few each week. Downloading, manual re-formatting, and parsing 60 submissions (including our own test submissions) took 10 to 60 minutes each week. A stricter policy saving more time would be to validate and automatically reject incorrect submissions.

AI Grading and Feedback. Depending on the number of questions in an assignment (e.g., 48 in assignment 1 and only 14 in the exam), prompting GPT-4o cost between 15 and 20 USD, and took about one to five minutes per submission (up to six hours for 60 submissions). This meant that students received preliminary grades and feedback within hours of the assignment deadlines. In contrast, human grading cost about 500 USD per assignment and took about 16 to 20 hours. With weekends and sick leave, students waited 3 to 12 days for their human-generated grades. They received no textual feedback from human

[1] Specifically *gpt-4o-2024-05-13*.

graders. Thus, LLM-generated grades and feedback can quickly become profitable in terms of time and money, especially for assignments that already have detailed rubrics. However, this conclusion only holds if LLM-grades are accurate.

3.2 Alignment Between Human and LLM Grading

Grades awarded by GPT-4o and the human graders were positively correlated for all four assignments and the exam (r = .815, .570, .768, .600, .866). However, humans frequently awarded higher grades (see Fig. 1).

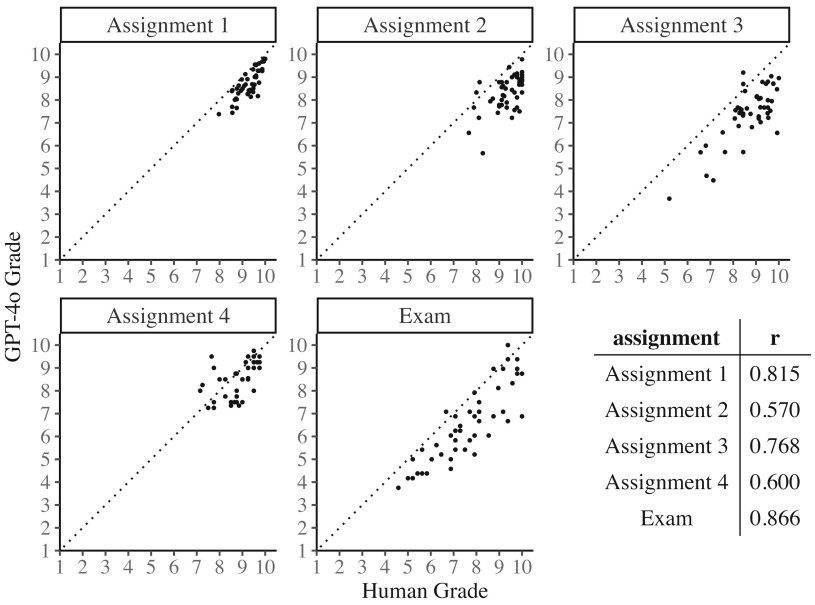


Fig. 1. Positive correlations, but humans awarded higher grades.

We also looked at the scores for 6345 question attempts (47 students x 135 questions). In 80.33% (5097) of cases, the human graders and GPT-4o assigned the same score; in 16.94% (1075), humans awarded more points; in 2.73% (173), GPT-4o awarded more points. We started investigating the 1248 disagreements by reading the explanations by GPT-4o and the human graders. At the time of writing we reviewed 210 of them and made the following discoveries.

Not All Disagreements Are LLM Mistakes. In 62 cases (29.52% of disagreements), the human made a mistake (e.g., missing a syntax error) and GPT-4o gave the correct score. In 42 disagreements (20%), both scores seem justifiable, with GPT-4o often being stricter than the human grader in their interpretation

of rubrics (e.g., penalizing typos in required strings). Put together, the LLM score can be considered correct for 49.52% of the 210 investigated disagreements. In seven cases (3.3%), both the human and LLM made mistakes. In 99 cases (47.14% of the 210 investigated disagreements), the human gave the correct score, and GPT-4o made a clear error.

Next, we look only at the 106 disagreements where the LLM made mistakes.

GPT-4o Sometimes Deviated from the Rubrics. We found 34 cases (32.08% of LLM mistakes) where GPT-4o deducted points for minor deviations from the example solution or for errors not covered by the rubrics, likely because it flagged them in its step-by-step "thoughts." This suggests that the example solution and model reasoning can sometimes introduce errors. Explicitly instructing the LLM to reference the rubrics in its "thoughts", may help prevent these mistakes.

Unfamiliar Code Led to False-Negatives and False-Positives. In 32 cases (30.19% of LLM mistakes), the LLM flagged correct solutions as incorrect or vice versa. In these cases, the students code was often messy, convoluted, or used unconventional solutions or uncommon packages. These are the most worrying mistakes GPT-4o made because there is no easy way of preventing them.

GPT-4o Lacks Certain Skills. In 17 cases (16.04% of LLM mistakes), GPT-4o failed to correctly count code lines, perform math, or evaluate the informativeness of plots. This explains the comparatively low correlation for assignment 2 ($r = .570$) which concerned plotting; and assignment 4 ($r = .600$) for which 20% of the grade were based on the number of code lines. For such tasks, LLM grading should likely be avoided or supplemented with unit tests or screenshots.

GPT-4o Interpreted Rubrics Too Literally. In 16 cases (15.09%, of LLM mistakes), GPT-4o penalized students for not using exact phrases (e.g., student wrote "you can't divide something by 0" instead of "division by zero"). This suggests that rubrics must avoid overly narrow language.

Our Mistakes Confused GPT-4o. In five cases (4.72% of LLM mistakes), we provided incorrect example solutions or introduced question dependencies without providing the necessary code. Human graders recognized that we made these mistakes and graded accordingly. In contrast, the LLM-scoring was correct based on the provided resources, but incorrect given the broader context.

GPT-4o Sometimes Overlooked Correct Comments. In two cases (1.89% of LLM mistakes) GPT-4o did not recognize comments added behind code instead of in a new line. This may be reconciled by explicitly prompting the LLM to look for such comments or by instructing students to avoid them.

What This Meant for Students. For some students, grades awarded by humans and GPT-4o were very similar or even identical. However, there were also those who would have received a failing grade (5/10) from the LLM and a good grade (8/10) from a human grader (see Fig. 1). So what did students think about this?

3.3 Reception by Students

We analyzed answers by 30 participants who gave their opinion on the preliminary grades and feedback. Most (23, 76.67%) liked the LLM-generated feedback and a quarter (8, 26.67%) specifically appreciated the optional coding challenge. Some (4, 13.33%) mentioned that they would have liked to receive feedback on every assignment (instead of only two). Students described feedback as "helpful" or "useful" and appreciated that it was "fast", "detailed", "specific", highlighted coding mistakes and suggested concrete improvements. However, some found the feedback "confusing", "difficult to read", "too long", or "pedantic". Ten participants (33.33%) indicated that they liked receiving timely preliminary grades. However, nine of those ten students (30%) disliked that they were often lower than their actual grades and four of them (13.33%) stated that this made them anxious or upset. Some mentioned that they would prefer question-wise scores and explanations instead of overall grades. Based on this analysis, students seem to value receiving timely grades and feedback. However, we aim to improve their presentation and accuracy for next years course.

For privacy reasons we do not share student submissions. However, we include our own example submissions and corresponding LLM-generated grades and feedback here: https://github.com/lukekorthals/canvas-llm-integration.

4 Main Takeaways and Recommendations

We created a reusable pipeline for AI grading and feedback of student assignments. In January 2025 we piloted the pipeline in an introductory programming course and found LLM grading and feedback to be fast and cost efficient. Overall, GPT-4o awarded relatively accurate grades (80.33% perfect agreement with human scores at the question level), but frequently and sometimes erroneously awarded lower grades than human graders. However, GPT-4o also spotted mistakes overlooked by humans. Students appreciated timely feedback and grades, but some reported that lower LLM-grades made them anxious. Importantly, we think that many of the LLM-scoring mistakes might be prevented by improving the tasks, rubrics, and instructions. Based on our analysis of disagreements, we specifically recommend others wanting to experiment with AI-grading to:

1. Write detailed subtractive rubrics (e.g., "subtract 0.5 points if X").
2. Avoid rubrics that require counting or other skills that LLMs lack.
3. Instruct the LLM to strictly adhere to the rubrics.
4. Avoid overly narrow language (e.g., "somehow explained X" rather than "explicitly mentioned X").

5. Provide your students not only with LLM-grades but also its reasoning.

We are currently validating the efficacy of these recommendations and investigate their generalizability in a broader sample of courses. We are also looking into open-source LLMs which make it easier to assess potential ethical concerns of AI-grading, including bias or the potential for jail-breaking.

Here https://github.com/lukekorthals/canvas-llm-integration you can find our code, and the assignments, rubrics, and prompts we used. We think it is a good starting point for anybody wanting to experiment with AI-grading, particularly Canvas-users and invite you to try it out.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Canvas. https://www.instructure.com/canvas
2. OpenAI Platform. https://platform.openai.com
3. TLC-FMG AI pilot project proposals. https://tlc.uva.nl/en/article/tlc-fmg-ai-pilot-project-proposals/
4. TLC FMG homepage. https://tlc.uva.nl/en/article-faculty/fmg-en/
5. Carless, D.: Learning-oriented assessment: conceptual bases and practical implications. Innov. Educ. Teach. Int. **44**(1), 57–66 (2007). https://doi.org/10.1080/14703290601081332
6. Dai, W., et al.: Can Large language models provide feedback to students? a case study on ChatGPT. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 323–325 (2023).https://doi.org/10.1109/ICALT58122.2023.00100, https://ieeexplore.ieee.org/document/10260740, iSSN: 2161-377X
7. Flodén, J.: Grading exams using large language models: a comparison between human and AI grading of exams in higher education using ChatGPT. Br. Edu. Res. J. **51**(1), 201–224 (2025). https://doi.org/10.1002/berj.4069
8. Golchin, S., Garuda, N., Impey, C., Wenger, M.: Large language models as MOOCs graders. arXiv:2402.03776 (2024)
9. Grévisse, C.: LLM-based automatic short answer grading in undergraduate medical education. BMC Med. Educ. **24**(1), 1060 (2024). https://doi.org/10.1186/s12909-024-06026-5
10. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007). https://doi.org/10.3102/003465430298487
11. Ishida, T., Liu, T., Wang, H., Cheung, W.K.: Large language models as partners in student essay evaluation. arXiv:2405.18632 (2024)
12. Jia, Q., Cui, J., Du, H., Rashid, P., Xi, R., Li, R., Gehringer, E.: LLM-generated feedback in real classes and beyond: perspectives from students and instructors, pp. 862–867 (2024).https://doi.org/10.5281/zenodo.12729974a, https://educationaldatamining.org/edm2024/proceedings/2024.EDM-posters.99/
13. Lynch, R., Hennessy, J.: Learning to earn? The role of performance grades in higher education. Stud. High. Educ. **42**(9), 1750–1763 (2017). https://doi.org/10.1080/03075079.2015.1124850

14. McMorran, C., Ragupathi, K., Luo, S.: Assessment and learning without grades? Motivations and concerns with implementing gradeless learning in higher education. Assess. Eval. High. Educ. **42**(3), 361–377 (2017). https://doi.org/10.1080/02602938.2015.1114584
15. Meyer, J., et al.: Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Comput. Educ. Artif. Intell. **6**, 100199 (2024). https://doi.org/10.1016/j.caeai.2023.100199
16. Morris, R., Perry, T., Wardle, L.: Formative assessment and feedback for learning in higher education: a systematic review. Rev. Educ. **9**(3), e3292 (2021). https://doi.org/10.1002/rev3.3292
17. OpenAI, et al.: GPT-4o system card. arXiv:2410.21276 (2024)
18. Slack: Slack. https://slack.com
19. Van Rossum, G., Drake Jr, F.L.: Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam (1995)
20. Voelkel, S., Varga-Atkins, T., Mello, L.V.: Students tell us what good written feedback looks like. FEBS Open Bio **10**(5), 692–706 (2020). https://doi.org/10.1002/2211-5463.12841
21. Williams, A.: Delivering effective student feedback in higher education: an evaluation of the challenges and best practice. Int. J. Res. Educ. Sci. **10**(2), 473–501 (2024). https://doi.org/10.46328/ijres.3404, https://ijres.net/index.php/ijres/article/view/3404
22. Winstone, N.E., Boud, D.: The need to disentangle assessment and feedback in higher education. Stud. High. Educ. **47**(3), 656–667 (2022). https://doi.org/10.1080/03075079.2020.1779687

Short-Term Gains, Long-Term Gaps: The Impact of GenAI and Search Technologies on Retention

Mahir Akgun[1(✉)] and Sacip Toker[2]

[1] College of Information Sciences and Technology, Pennsylvania State University, University Park 16802, USA
makgun@psu.edu

[2] Information Systems Engineering, Atilim University, Ankara, Türkiye
sacip.toker@atilim.edu.tr

Abstract. The rise of Generative AI (GenAI) tools, such as ChatGPT, has transformed how students access and engage with information, raising questions about their impact on learning outcomes and retention. This study investigates how GenAI (ChatGPT), search engines (Google), and e-textbooks influence student performance across tasks of varying cognitive complexity, based on Bloom's Taxonomy. Using a sample of 123 students, we examined performance in three tasks: [1] knowing and understanding, [2] applying, and [3] synthesizing, evaluating, and creating. Results indicate that ChatGPT and Google groups outperformed the control group in immediate assessments for lower-order cognitive tasks, benefiting from quick access to structured information. However, their advantage diminished over time, with retention test scores aligning with those of the e-textbook group. For higher-order cognitive tasks, no significant differences were observed among groups, with the control group demonstrating the highest retention. These findings suggest that while AI-driven tools facilitate immediate performance, they do not inherently reinforce long-term retention unless supported by structured learning strategies. The study highlights the need for balanced technology integration in education, ensuring that AI tools are paired with pedagogical approaches that promote deep cognitive engagement and knowledge retention.

Keywords: AI in Education · Generative AI · Search Tools · Retention · Bloom's Taxonomy

1 Introduction

Generative AI (GenAI) tools offer the ability to synthesize information from various sources, enabling students to write essays and apply knowledge in new contexts with unprecedented ease. This represents a significant shift from the previous reliance on search engines like Google, where students manually gathered and integrated information, to a more streamlined and automated process facilitated by GenAI.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 44–52, 2025.
https://doi.org/10.1007/978-3-031-99264-3_6

Traditional search engines support learning by providing access to diverse information sources, requiring students to actively analyze and synthesize data–a process well-suited to higher levels of Bloom's taxonomy, such as analysis and creation. In contrast, GenAI tools streamline this process by delivering synthesized, ready-to-use insights, reducing the cognitive effort needed to locate and integrate information. While this efficiency may benefit short-term performance, it raises important questions about the long-term impact on learning, particularly across tasks of varying cognitive complexity. This study explores the role of different technologies–GenAI tools such as ChatGPT, search engines like Google, and traditional textbooks–in influencing students' achievement across learning tasks of varying complexities.

1.1 Search Systems and Learning

Modern search systems have evolved into rich learning environments where users actively acquire knowledge through interaction with online content [7]. Prior research highlights that advanced learners typically gain more from search-based learning than novices [11,14]. Moreover, combining search with structured instruction–such as video lectures–tends to produce stronger learning outcomes than using either in isolation [12].

1.2 Learning Through Search: A Taxonomical Approach

Bloom's taxonomy provides a widely used framework for classifying cognitive processes in learning, ranging from basic recall to creation [1]. Vakkari suggested that Bloom's taxonomy can be applied to common search tasks involving querying to find topically relevant information or sources [15]. These tasks include formulating searches and evaluating results (source selection). Searchers require factual knowledge to choose search terms and identify relevant sources from result lists. They also utilize procedural knowledge to execute the search process. The cognitive processes involved in these tasks primarily include remembering–recognizing search terms from the task and results, recalling them from memory, and identifying relevant items in the result list. The taxonomy has been used to characterize learning requirements in search tasks (e.g., [2,9]). Jansen et al. [9] found that mid-level cognitive tasks such as applying and analyzing required the most search effort, while both lower- and higher-level tasks elicited similar, often simplistic search behaviors–suggesting reliance on internal knowledge for complex tasks. Similarly, Arguello et al. [2] showed that increasing task complexity leads to greater user interaction, especially with rich content like images and videos.

1.3 Present Study

Several studies highlight the challenges search engines face in comprehending students' information needs due to their use of long, natural language phrases

and ambiguous queries (e.g., [4,5,10]). Despite recent developments to address this issue in information retrieval, formulating queries to access information has remained a challenge for students. This issue can be addressed using GenAI tools (e.g., [16]). The emergence of GenAI tools such as ChatGPT has led learners to increasingly rely on these conversational agents (e.g., [8]). This transformation prompts critical inquiries about the role of AI tools in the learning process and their potential impact on memory and retention. The increasing use of GenAI tools like ChatGPT in education has raised questions about their influence on learning outcomes, particularly in comparison to traditional tools such as search engines and e-textbooks. While GenAI excels in interpreting natural language and providing synthesized responses, its impact on memory retention and deeper cognitive engagement remains unclear. Unlike traditional tools that require learners to search, analyze, and synthesize information manually, GenAI potentially reduces cognitive effort–raising concerns about its effects on long-term learning.

This study addresses this gap by comparing how GenAI (ChatGPT), search engines (Google), and e-textbooks affect student performance across tasks aligned with Bloom's taxonomy. It investigates both short-term achievement and long-term retention across tasks requiring lower- to higher-order cognitive skills. We address the following research questions:

- How does the use of different tools (e-textbook, Google, ChatGPT, or none) affect student achievement across tasks of varying cognitive complexity?
- How do these tools influence students' retention of learning outcomes across different cognitive levels?

2 Method

2.1 Participants

152 students from an information systems department at a private university agreed to participate in this study. However, data from 29 students who did not complete all tasks as instructed were excluded. Consequently, the final sample consisted of 123 students. Participants completed two pretests and two self-efficacy scales to measure their knowledge of data privacy (M=54.81, SD = 27.10 out of 100), familiarity with ChatGPT (M=71.29, SD = 9.32 out of 100), search self-efficacy (M= 49.52, SD = 9.22 out of 70), and search experiences.

2.2 Research Design

This study employed an experimental pre-and post-test design with random assignment. This design included a control and three experimental groups, each exposed to different tools. Initially, participants were randomly assigned to one of four groups: control, e-textbook, Google, and ChatGPT access. The e-textbook group had access to a PDF version of the course textbook on data privacy,

with no other resources allowed. The Google group used only the Google search engine, and the ChatGPT group relied solely on ChatGPT. The control group had no access to these tools during the experiment. The experiment was conducted in a laboratory environment under the supervision of the course instructor and teaching assistants. Participants were continuously monitored. Three tasks were designed based on Bloom's Revised Taxonomy [1], progressing from lowest to highest complexity: Task 1 (remembering/understanding), Task 2 (applying), and Task 3 (analyzing/evaluating/creating). All participants completed the tasks sequentially: Task 1, Task 2, and Task 3 (see Table 1).

Table 1. Task Descriptions Across Bloom's Taxonomy Levels

Task 1: Remembering & Understanding	Task 2: Applying	Task 3: Analyzing, Evaluating & Creating
Please answer the following questions and submit them as a Word file to the Moodle Course site: (1)What is the EU General Data Protection Regulation (GDPR)? List GDPR requirements for processing personal information. (2) What types of information are protected by the GDPR? (3)What are the main differences and similarities between privacy security and ethics?	Suppose that you work for a company as a privacy and security expert with proficiency on legal and technical aspects of GDPR. The company's management wants to make sure it has policies and procedures in place to protect the privacy of visitors to its website. You have been asked to provide advice to the company about what needs to be included in the policies to ensure that they comply with the EU GDPR.	Examine the Optimum Customer Privacy Notice and evaluate whether it complies with the EU GDPR. Base your evaluation on the GDPR requirements. After the evaluation identify gaps in the privacy policy. What are your suggestions for improving the privacy policy (PR)? Please prepare a list of suggestions to improve the PR document.

2.3 Group Formation

Participants were randomly assigned to four groups: control, e-textbook, Google, and ChatGPT. To ensure balanced distribution of prior knowledge and experience, a two-step cluster analysis was used based on factors such as data privacy knowledge, search self-efficacy, and familiarity with AI tools. Participants from each cluster were evenly distributed across the groups, maintaining diversity in expertise and ensuring comparable conditions. Gender distribution was also balanced, with the final group sizes as follows: control (n = 31), e-textbook (n = 25), Google (n = 29), and ChatGPT (n = 38).

2.4 Procedure

Groups completed three learning activities in the study. Each learning activity consisted of two components: (1) a task completed during the lab session, where students used their assigned tools (e-textbook, Google, ChatGPT, and none) to work through scenario-based exercises, and (2) a quiz administered in Moodle immediately after the task, featuring both multiple-choice and open-ended questions designed to assess individual comprehension without access to external tools.

For each task, one lecture/demonstration session (50 min) and one lab session (100 min) were allocated in the course. The lecture/demonstration session took place before the lab session and was used to review relevant course content, introduce key concepts, and demonstrate how they could be applied to similar tasks. In the lab sessions, students completed their assigned tasks with full access to their designated tools (e-textbook, Google, or ChatGPT) to support their engagement with the scenario-based exercises. However, during the quiz phase, all students completed the Moodle-based assessment independently, without access to any external tools. The quizzes evaluated students' grasp of key concepts and their performance on questions requiring the cognitive skills targeted in the corresponding task.

Three weeks after completing each task, students took a post-test designed to assess long-term retention of key concepts. The post-test included multiple-choice and short-answer questions that were conceptually aligned with the original task assessments. Figure 1 illustrates the procedure followed in the study.

2.5 Data Analysis and Measures

For the first research question, we collected assessment scores for three progressive tasks, which served as the dependent variables. The independent variable was the group category, consisting of four groups: three experimental groups and one control group. MANOVA analyses were performed [6]. After a significant result, we followed up univariate ANOVAs for each task assessment. We checked the assumptions for both MANOVA and univariate ANOVA. For the second research question, we used repeated-measures ANOVA. This was appropriate because assessment and retention scores were measured on two different occasions for the same groups [6]. We ensured the content validity of the assessments by aligning them with the goals of the tasks and the specifics of the privacy topics.

3 Results

3.1 Results for Research Question 1

There were significant differences in task-level assessment scores based on the groups. The multivariate analysis indicated a significant impact of groups on the three task-level assessment scores, with Wilk's $\lambda = .677$, $F(9, 284.90) = 5.513$, $p < .001$, and partial $\eta^2 = .122$, explaining 12.2% of the total variance. Univariate ANOVA findings were examined.

Task 1 Assessment Scores. The univariate ANOVA results for Task 1 showed significant differences between groups, $F(3, 119) = 10.650$, $p < .001$, partial $\eta^2 = .212$. The ChatGPT group outperformed the E-textbook and control groups significantly, with mean differences of 16.60 and 22.39, respectively ($p < .01$). The Google group also scored significantly higher than the control group, with a mean difference of 18.90 ($p < .01$). The ChatGPT group demonstrated the highest performance (see Fig. 2).

Task 2 Assessment Scores. For Task 2, the univariate ANOVA results showed significant differences between groups, $F(3, 119) = 7.279$, $p < .001$, partial $\eta^2 = .155$. The control group scored significantly lower than both the Google and ChatGPT groups, with mean differences of 18.81 and 20.15, respectively ($p < .01$). The ChatGPT group again showed the best performance (see Fig. 2).

Task 3 Assessment Scores. For Task 3, no significant differences between groups were found, $F(3, 119) = .768$, $p = .514$. The results indicated that the control group had the best performance, while the E-textbook, Google, and ChatGPT groups had lower scores compared to the control group (see Fig. 2).

3.2 Results for Research Question 2

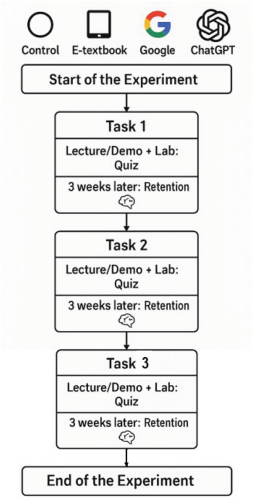


Fig. 1. Procedure followed in the study

Descriptive statistics for retention tests are provided in Table 2. The statistical analysis results of the study reveal important insights into the performance and retention of different groups across three tasks.

Task 1 Results. A significant interaction between assessment and retention scores was found (Wilk's $\lambda = .885$, $F(3, 119) = 5.175$, $p < .01$, partial $\eta^2 = .115$), indicating that score changes differed across groups. The ChatGPT and Google groups showed notable retention declines despite high initial performance, while control and e-textbook groups demonstrated more stable performance over time.

Task 2 Results. None of the effects produced significant results. Although there were decreases in performance for the ChatGPT and Google groups and increases for the E-textbook and control groups, these changes were not statistically significant.

Task 3 Results. Although all groups showed performance decreases from assessment to retention, these changes were not significant.

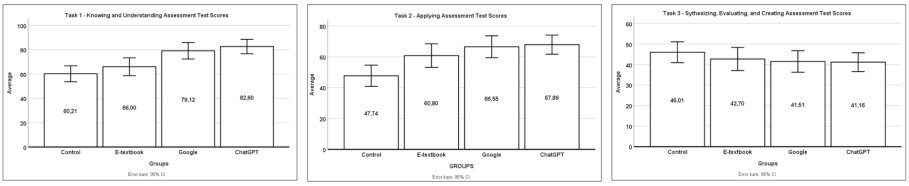


Fig. 2. Assessment Scores

4 Discussion

4.1 Impact on Immediate Performance

The results indicate that students using ChatGPT and Google performed significantly better than the control group on lower- and mid-level cognitive tasks (Tasks 1 and 2). These results align with Vakkari's findings on the relationship between Bloom's taxonomy and search tasks [15]. These tools provided quick access to structured information, which likely reduced cognitive load and helped students complete tasks more efficiently. However, this advantage did not extend to higher-order tasks (Task 3), where the need for critical thinking, synthesis, and creation diminished the benefits of passive information retrieval. As tasks became more complex, the capability to retrieve well-structured responses was no longer sufficient. These tasks demand active cognitive engagement beyond factual recall, which may not have been fully supported by the way students interacted with AI tools and search engines.

Table 2. Descriptive information about retention tests

Group	Task 1		Task 2		Task 3	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Control	59.3	22.06	45.9	20.0	44.2	16.3
E-textbook	67.6	24.94	61.3	22.1	41.9	14.9
Google	67.8	25.4	63.1	20.1	40.6	13.8
ChatGPT	65.5	22.1	64.5	21.1	39.9	15.1

4.2 Impact on Retention

While ChatGPT and Google provided immediate performance advantages, these benefits did not translate into long-term retention. In retention tests three weeks later, ChatGPT and Google groups' performance declined to align with the e-textbook group, consistent with natural forgetting processes. These tools facilitate efficient information access but do not inherently strengthen memory pathways for sustained retention.

For the applying task, performance decreases were not statistically significant, suggesting no tool provided superior retention. As task complexity increased, retention differences across groups became less pronounced. In higher-order tasks, the control group showed highest retention, while the e-textbook group exhibited smaller decreases, suggesting that reading and comprehending textual information reinforced retention more effectively than retrieving pre-structured responses, aligning with Purdy's findings [13] on e-textbooks fostering deeper scholarly engagement.

The relatively lower retention in the GenAI and Google groups may be due to the manner in which students interacted with these tools. While providing quick access to synthesized information, these technologies do not inherently reinforce knowledge integration unless paired with learning strategies that engage memory consolidation. ChatGPT sometimes generated generic responses lacking depth, limiting retention effectiveness for complex tasks. Similarly, effective use of Google required deeper query formulation and critical filtering of information, which may have introduced variability in retention outcomes [11]. Ultimately, ensuring long-term retention requires instructional approaches beyond initial task completion. Given natural forgetting processes [3], additional reinforcement activities–such as reflective exercises, spaced repetition, and active recall–may be necessary to solidify learning regardless of technology used.

4.3 Limitations and Future Research Directions

This study has limitations that should be noted. First, the sequential task order may have introduced cumulative advantages or disadvantages, particularly for AI and search engine users, making it difficult to isolate individual task effects. Future research should counterbalance task order to mitigate sequencing effects. Second, the sample of 123 students from a single institution may limit generalizability across diverse educational contexts. Expanding research to larger, more diverse participant pools would help validate these findings.

References

1. Anderson, L.W., Krathwohl, D.R.: A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Complete Edition. Addison Wesley Longman, Inc. (2001)
2. Arguello, J., Wu, W.C., Kelly, D., Edwards, A.: Task complexity, vertical display and user interaction in aggregated search. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 435–444 (2012)
3. Baddeley, A.D.: Human Memory: Theory and Practice. Psychology Press (1997)
4. Bilal, D.: Children's use of the yahooligans! web search engine: I. cognitive, physical, and affective behaviors on fact-based search tasks. J. Am. Soc. Inf. Sci. **51**(7), 646–665 (2000)
5. Druin, A., Foss, E., Hutchinson, H., Golub, E., Hatley, L.: Children's roles using keyword search interfaces at home. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 413–422 (2010)

6. Field, A.: Discovering Statistics Using IBM SPSS Statistics. Sage, London, UK (2013)
7. Hansen, P., Rieh, S.Y.: Recent advances on searching as learning: an introduction to the special issue. J. Inf. Sci. **42**(1), 3–6 (2016)
8. Jalon, J.B., Chua, G.A., de Luna Torres, M.: ChatGPT as a learning assistant: its impact on students learning and experiences. Int. J. Educ. Math. Sci. Technol. **12**(6), 1603–1619 (2024)
9. Jansen, B.J., Booth, D., Smith, B.: Using the taxonomy of cognitive learning to model online searching. Inf. Process. Manage. **45**(6), 643–663 (2009)
10. Jochmann-Mannak, H., Huibers, T.W., Lentz, L., Sanders, T.: Children searching information on the internet: performance on children's interfaces compared to google. In: Workshop on Accessible Search Systems 2010 (2010)
11. Lu, Y., Hsiao, I.H.: Personalized information seeking assistant (PiSA): from programming information seeking to learning. Inf. Retrieval J. **20**, 433–455 (2017)
12. Moraes, F., Putra, S.R., Hauff, C.: Contrasting search as a learning activity with instructor-designed learning. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 167–176 (2018)
13. Purdy, J.P.: Why first-year college students select online research resources as their favorite. First Monday (2012)
14. Roy, N., Moraes, F., Hauff, C.: Exploring users' learning gains within search sessions. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, pp. 432–436 (2020)
15. Vakkari, P.: Searching as learning: a systematization based on literature. J. Inf. Sci. **42**(1), 7–18 (2016)
16. Zhang, M., Yang, X.: Google or ChatGPT: who is the better helper for university students. Educ. Inf. Technol. 1–22 (2024)

Enhancing Teacher Support in Learning Technologies: A Human-Centered Approach with WearableLearning

Allison Poh[1](✉), Yuqian Shi[2], Francisco Enrique Vicente Castro[3], and Ivon Arroyo[1]

[1] University of Massachusetts Amherst, Amherst, MA 01002, USA
apoh@cs.umass.edu
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] New York University, New York, NY 10012, USA

Abstract. Learning technologies are widely used in K-12 classrooms but often fall short in meeting teachers' needs. This study employs a multi-method, human-centered approach to design teacher support for WearableLearning (WL), a learning technology for math and computational thinking through embodied learning. We began with a survey of 52 K-12 teachers to assess how the learning technologies they use in their classrooms support or fail to support them. We identified eight areas where learning technologies offer support and five areas where they lack support, and used these insights to inform the design of a preliminary teacher dashboard for WL. We conducted two rounds of interviews with three teachers to refine the design through iterative feedback. Finally, we administered a questionnaire to eight pre-service teachers to further explore teacher needs. This process resulted in a low-fidelity teacher dashboard prototype, which will undergo cognitive walkthroughs in future work. The findings from our study can be used to better design teacher support in learning technologies.

Keywords: Teacher dashboards · learning technologies · human-centered design

1 Introduction

Learning technologies are playing an increasingly central role in K-12 education, promising to enhance both student learning and teacher instruction [1]. However, while these technologies aim to improve education, teachers often struggle to use them due to insufficient training and understanding of how to apply them effectively [2]. Furthermore, many technologies fail to adequately support teachers in monitoring students and understanding students' processes, misconceptions, and knowledge gaps [3].

In recent years, there has been growing interest in enhancing teacher support through dashboards [3, 4]. Teacher dashboards empower teachers with real-time student data, enabling deeper insights, reflection, and data-driven decision-making [3, 5]. While prior work has focused on teacher dashboards for traditional learning technologies (*i.e.,* systems where students remain seated at desks) [3–7], teacher dashboards for less traditional

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 53–60, 2025.
https://doi.org/10.1007/978-3-031-99264-3_7

technologies that encourage physical movement remain underexplored. One such system is WearableLearning (WL) [8, 9], which promotes learning through embodied gameplay, blending the physical classroom with the digital space.

While prior work has focused on professional development (PD) to support teachers adopting WL [10], no work has explored supporting teachers in real-time while students play WL games. Questions remain regarding how teachers can monitor students and view their progress and activity. To address this gap, we used a multi-method, human-centered approach to build teacher support for WL. In doing so, we address the following research questions: *How do learning technologies support or fail to support teachers?* and *How can teacher support be integrated into WL to best meet teachers' needs?*

2 Background

2.1 Challenges of Learning Technologies in the Classroom

Learning technology is a broad term that encompasses any instructional tools designed to aid student learning [2, 11]. The pace at which these technologies are developed often outpaces the ability of teachers to learn how to use them effectively [2]. This is partly due to the Gartner hype cycle, which describes how new technologies surge in adoption due to excitement but decline when expectations aren't met [2, 12, 13]. When teachers try to incorporate new technologies but encounter challenges early on, they may abandon the technology and become apprehensive about trying new ones [2].

One way to address these challenges is through PD [14, 15]. Ertmer and Ottenbreit-Leftwich (2010) argue to support "21st-century teaching and learning, we need to help teachers understand how to use technology to facilitate meaningful learning." While PD can help teachers, for it to be effective, it must demonstrate the technology is easy to use and beneficial for instruction, which is often not achieved in PD models [15].

PD may help teachers adopt a technology, but supporting teachers *while* students use it is another challenge. Teachers need real-time, actionable insights to support their awareness, reflection, and decision-making [3, 16]. More specifically, in a study on real-time learning analytics, Holstein *et al.* (2017) identified seven needs of teachers from learning technologies: (1) seeing students' thought processes, (2) identifying who is stuck or (3) "almost there," (4) supporting multiple students simultaneously, (5) monitoring misbehavior, (6) detecting misconceptions, and (7) recognizing patterns of careless errors [3]. These needs are often met through real-time teacher dashboards [4].

2.2 Dashboards for Supporting Teachers

Dashboards empower teachers with real-time student data, enabling deeper insights, reflection, and data-driven decision-making [3, 5]. Specifically, Poh *et al.* (2023) identified seven key tasks supported by teacher dashboards: (1) viewing data at student, group, and class levels, (2) viewing students' task performance, (3) providing feedback and support to students, (4) personalizing learning, (5) monitoring and supporting collaboration, (6) orchestrating activities, and (7) reflecting after activities.

Teacher dashboards support these tasks through simple visualizations, symbols, colors, and the ability to adjust data granularity [4–7]. As Molenaar and Knoop-van Campen

(2017) noted in regard to the *distributed cognition theory*, "instruments can indeed support professionals, when these instruments fit seamlessly into the activities of a professional" [5, 17]. That is, effective teacher dashboards must align with teachers' tasks to provide meaningful support, and features like simple visualizations and colors best align with the primary task of monitoring and viewing student data [4, 5].

AI integration in educational technologies has grown in recent years [18–22]. As teacher dashboards become increasingly AI-powered, questions remain about teachers' willingness to adopt and trust AI. Through discussions with teachers, we explore their perspectives on AI support, discussed in *Sect.* 4: *Results and Discussion.*

2.3 WearableLearning

WL is a learning technology that promotes math and computational thinking through *game creation* and *gameplay* [8, 9]. Unlike traditional technologies that confine students to desks, WL creates a blended learning environment by integrating physical and digital spaces, supporting active learning through movement [9, 10, 23].

Games in WL are created using a simplified drag-and-drop programming language based on finite state machines (FSM). An FSM is an abstract model that represents a system's behavior (in this case, a game) through two main components: *states* and *transitions*. While prior research has focused on how students create games [9, 23, 24], this paper focuses on game playing, specifically on supporting teachers while students play.

One example of an embodied WL game is *Tangrams Race*, a geometric relay game where students race across the classroom to collect tangram pieces based on clues displayed on their armband-worn devices. Other examples include stepping forward and backward on a large number line, scavenger hunts, and dance-based activities.

Compared to traditional learning technologies, WL presents greater monitoring challenges since students are moving around and the smaller, wearable devices are harder to monitor. A key limitation of the current WL platform is the lack of effective teacher support during gameplay [24]. To address this gap, this study explores ways to better support teachers while students play games by first examining how other technologies provide support and then exploring how to integrate this into WL.

3 Methodology

To understand how to best support teachers within the WL platform, we used a multimethod, human-centered design approach with three phases.

3.1 Survey

We recruited teachers via publicly listed emails on school district websites. 52 teachers (11 primary, 16 middle, 25 high school) from diverse public and private schools in the Northeastern US participated. The survey included five open-ended questions (*e.g., What learning technologies support you and how?*). We analyzed responses with word clouds and affinity diagramming to identify areas of support and lack thereof.

3.2 Interviews

Based on the survey findings, we designed an initial mockup of a teacher dashboard for WL, which guided our semi-structured interviews. We recruited two middle school teachers from a prior WL PD study, neither of whom had participated in the survey. Additionally, we recruited a primary school teacher who was a member of the WL research team at the time but whose research focused on creating WL games rather than platform design or development. They had no apparent stake in a dashboard for WL.

The first interview round explored teachers' experiences with WL and their perspectives on how it supports them. These 45–60-min Zoom interviews were conducted individually and recorded with consent. While discussions were teacher-driven, we prepared guiding questions such as *In what ways does WL currently support/not support you?* In the second half, we presented our initial teacher dashboard mockup and asked for feedback, including *What data is missing that you wish you could see?* Using grounded theory, we gathered insights from our discussions to revise the mockup.

The second interview round focused on teachers' perspectives on the revised mockup. These 30-min Zoom interviews were conducted individually and recorded with consent. As in the first round, discussions were teacher-driven, with guiding questions like *What information do you see here?* and *How would you use this information in your classroom?* Grounded theory was used again to inform a low-fidelity prototype.

3.3 Questionnaire

In recognition of our small interview sample, we administered a questionnaire to eight pre-service teachers in a Master of Arts in Teaching program. Before responding, participants played *Tangrams Race*. The questionnaire reflected themes from the interviews, including student monitoring, a common concern among interviewees.

4 Results and Discussion

From the survey, teachers cited 108 learning technologies, with Google technologies (*e.g.,* Chromebooks, Google Classroom) being the most frequently mentioned. Through affinity diagramming, we identified eight areas where learning technologies *offer support* (see Table 1) and five areas they *lack support* (see Table 2) for teachers. While most teachers' needs identified by Holstein *et al.* (2017) are addressed to some capacity by the *offer support* areas (*e.g.,* through assessment reports), an alarming issue is the lack of support for the need of student monitoring. Based on our findings, we identified two priority areas for our mockup: *assessment and feedback* and *student monitoring*.

The top of Fig. 1 shows two key views from our initial dashboard mockup: *Class* and *Individual*. The *Class* view displays overall student progress, including each student's placement and performance on a simplified FSM of the game (*e.g.,* red circle represents a struggling student). The *Individual* view shows more detailed performance data for a single student, along with AI-generated suggestions for assisting them. Together, these views aim to support teachers in viewing both class and student data (*e.g.,* task performance) and real-time support and monitoring, addressing three of the seven key teacher tasks identified by Poh *et al.* (2023). (Although not shown in Fig. 1, our mockup also includes *Orchestrate*, *Task*, and *Report* views for the remaining tasks.)

Table 1. Areas where learning technologies *offer support* to teachers.

Area	Sample Response
Instruction and Content Delivery	"I use [it] to get quick reports that I can use to guide instruction."
Personalized Learning	"[It] allows me to assign and differentiate work for students."
Engagement and Interactive Learning	"[They] allow me to be able to post interactive websites or work online."
Assessment and Feedback	"They allow me to see student growth, performance data, and student choice instantaneously."
Classroom Organization	"[It] definitely supports me by allowing me to keep my classroom and gradebook much more organized."
Communication and Collaboration	"[It] allows me to tutor students if they can't be in school, talk to parents who can't come in for meetings, or teach virtually."
Efficiency and Time Management	"Anything that can simplify the work that I have to do behind the scenes, improves the work that my students see."
Technology Integration	"[They] support me as a teacher because I have the ability to access various platforms with my students."

Our first round of interviews revealed that our dashboard mockup needed refinement to better support teachers. Teachers emphasized the importance of seeing the *exact* screens of students. While they could view progress in the initial mockup, they couldn't always understand the exact challenges students faced (*e.g., Why hasn't a student progressed?*). Teachers also preferred the *exact* FSM over the simplified version. Initially, we simplified the FSM based on research suggesting that teachers prefer simple visualizations [4]. However, our findings highlighted an important counterpart to simplicity: *familiarity*. Teachers expressed that the exact FSM was simpler because they were familiar with it from creating games. This aligns with the *familiarity heuristic* in which people subjectively prefer what they recognize [25]. In response, we made two key changes to the mockup (bottom of Fig. 1): a live view of students' screens and an exact FSM. In the

Table 2. Areas where learning technologies *lack support* for teachers.

Area	Sample Response
Availability	"We need more [learning technologies] that are accessible and free and teachers do not have to constantly search for them."
Usability	"Some are so busy that it is hard to quickly do what you want to."
Customizability	"[M]ore customizable interfaces so that I can construct my own unique interactives and activities."
Student Monitoring	"I wish we could monitor more easily. We use [monitoring software] but its multiple steps and takes up the whole screen and then becomes another "tab" to monitor."
Professional Development	"I wish that I received more [PD] on using more platforms. I feel like there [are] more websites and platforms to use but finding the time to implement them with lack of training can be a challenge."

second round of interviews, teachers strongly expressed the benefits of these changes for monitoring. Pre-service teachers also echoed this, with one also suggesting a student device pause feature that we added to the low-fidelity prototype.

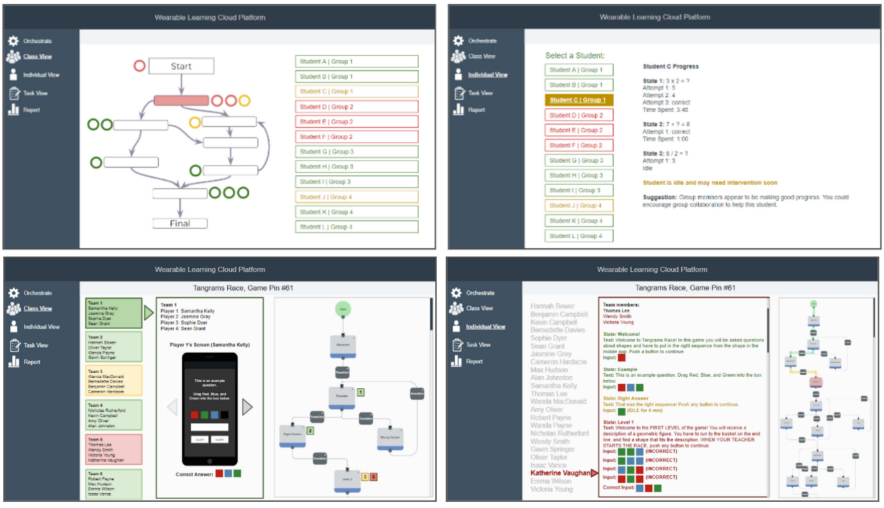


Fig. 1. Initial mockup *Class* (top-left) and *Individual* (top-right) views along with the revised *Class* (bottom-left) and *Individual* (bottom-right) views.

We also discussed AI-generated suggestions with interviewed teachers. In the first round, one teacher was apprehensive to use them over their own inferences, while the other found them too verbose. In response, we removed the AI suggestions, and in the second round, all teachers preferred seeing more student data over AI suggestions.

While our study suggests preliminary results towards teachers not having interest in AI-generated support at the moment, we wish to explore this more in the future.

5 Conclusion

In this study, we used a multi-method, human-centered approach to design a teacher dashboard. Our analysis identified eight areas where learning technologies offer support and five where they lack support for teachers. Additionally, our results suggest the importance of showing students' exact screens and designing with teacher familiarity in mind. Regarding AI support, our results suggest teachers have limited interest at the moment. Future work will evaluate our prototype through cognitive walkthroughs.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Kanuka, H.: Understanding e-learning technologies-in-practice through philosophies-in-practice. In: Anderson, T. (ed.) The Theory and Practice of Online Learning, 2nd edn., pp. 91–111. AU Press, Athabasca University (2008)
2. Lewis, G.: Learning Technology. Oxford University Press, Oxford (2017)
3. Holstein, K., McLaren, B. M., Aleven, V.: Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In: Proceedings of the 7th International Learning Analytics & Knowledge Conference, pp. 257–266 (2017)
4. Poh, A., Castro, F. E., Arroyo, I.: Design principles for teacher dashboards to support in-class learning. In: Proceedings of the 17th International Conference of the Learning Sciences, pp. 736–743 (2023)
5. Molenaar, I., Knoop-van Campen, C.: Teacher dashboards in practice: usage and impact. In: Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, pp. 125–138 (2017)
6. Xhakaj, F., Aleven, V., McLaren, B. M.: Effects of a teacher dashboard for an intelligent tutoring system on teacher knowledge, lesson planning, lessons and student learning. In: Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, pp. 315–329 (2017)
7. van Leeuwen, A., Rummel, N., van Gog, T.: What information should CSCL teacher dashboards provide to help teachers interpret CSCL situations? Int. J. Comput.-Support. Collab. Learn. **14**(3), 261–289 (2019). https://doi.org/10.1007/s11412-019-09299-x
8. WearableLearning Homepage. http://wearablelearning.org/. Accessed 7 May 2025
9. Arroyo, I., Closser, A.H., Castro, F., Smith, H., Ottmar, E., Micciolo, M.: The WearableLearning platform: a computational thinking tool supporting game design and active play. Technol. Knowl. Learn. **28**(4), 1815–1824 (2023)
10. Smith, H., Closser, A.H., Ottmar, E., Arroyo, I.: Developing math knowledge and computational thinking through game play and design: a professional development program. Contemp. Issues Technol. Teach. Educ. **20**(4), 660–686 (2020)
11. Leahy, S.M., Holland, C., Ward, F.: The digital frontier: Envisioning future technologies impact on the classroom. Futures **113**, 102422 (2019)

12. Prinsloo, T., Van Deventer, J. P.: Using the gartner hype cycle to evaluate the adoption of emerging technology trends in higher education—2013 to 2016. In: Emerging Technologies for Education: 2nd International Symposium SETE, pp. 49–57. Springer International Publishing, Cape Town (2017)
13. Fenn, G., Raskino, M.: Mastering the Hype Cycle: How to Choose the Right Innovation at the Right time. Gartner Inc., USA (2008)
14. Ertmer, P.A., Ottenbreit-Leftwich, A.T.: Teacher technology change: how knowledge, confidence, beliefs, and culture intersect. J. Res. Technol. Educ. **42**(3), 255–284 (2010)
15. Love, M.L., Simpson, L.A., Golloher, A., Gadus, B., Dorwin, J.: Professional development to increase teacher capacity for the use of new technologies. Interv. Sch. Clin. **56**(2), 115–118 (2020)
16. Jørnø, R.L., Gynther, K.: What constitutes an "actionable insight" in learning analytics? J. Learn. Anal. **5**(3), 198–221 (2018)
17. Hutchins, E.: Distributed cognition. In: International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science (2000)
18. Gobert, J. D., Dickler, R., Adair, A.: Using an AI-based dashboard to help teachers support students' learning progressions for science practices. In: Handbook of Research on Science Learning Progressions, pp. 488–498 (2024)
19. Aleven, V., Blankestijn, J., Lawrence, L., Nagashima, T., Taatgen, N.: A dashboard to support teachers during students' self-paced AI-supported problem-solving practice. In: European Conference on Technology Enhanced Learning, pp. 16–30 (2022)
20. Wiedbusch, M.: Intelligent teacher dashboards that support students' self-regulated learning, engagement, and teachers' decision making. Front. Educ. **6** (2021)
21. Lee-Cultura, S., Sharma, K., Giannakos, M.N.: Multimodal teacher dashboards: Challenges and opportunities of enhancing teacher insights through a case study. IEEE Trans. Learn. Technol. **17**, 181–201 (2023)
22. Kulkarni, A., Eagle, M.: Towards understanding the impact of real-time AI-powered educational dashboards (RAED) on providing guidance to instructors. arXiv preprint arXiv:2107.14414 (2021)
23. Rasul, I., Castro, F., Arroyo, I.: Towards embodied wearable intelligent tutoring systems. In: International Conference on Intelligent Tutoring Systems, pp. 298–306 (2023)
24. Arroyo, I., et al.: Aligning AIED systems to embodied cognition and learning theories. In: International Conference on Artificial Intelligence in Education, pp. 3–17 (2024)
25. Dasgupta, A.: Experts' familiarity versus optimality of visualization design: how familiarity affects perceived and objective task performance. Cogn. Biases Visualizat. 75–86 (2018)

SSRLBot: Designing and Developing a Large Language Model-Based Agent Using Socially Shared Regulated Learning

Xiaoshan Huang(✉), Jie Gao(✉), and Haolun Wu

McGill University, Montreal, QC, Canada
{xiaoshan.huang,jie.gao3,haolun.wu}@mail.mcgill.ca

Abstract. Large language model (LLM)–based agents have emerged as pivotal tools in assisting human experts across various fields by transforming complex tasks into more efficient workflows and providing actionable stakeholder insights. Despite their potential, the application of LLM-based agents for medical education remains underexplored. The study aims to assist in evaluating the students' process and outcomes on medical case diagnosis and discussion while incorporating the theoretical framework of Socially Shared Regulation of Learning (SSRL) to assess student performance. SSRL emphasizes metacognitive, cognitive, motivational, and emotional interactions, highlighting the collaborative management of learning processes to improve decision-making outcomes. Grounded in SSRL theory, this tool paper introduces SSRLBot, an LLM-based agent designed to enable team members to reflect on their diagnostic performance and the key SSRL skills that foster team success. SSRLBot's core functions include summarizing dialogue content, analyzing participants' SSRL skills, and evaluating students' diagnostic results. Meanwhile, we evaluated SSRLBot through diagnostic conversation data collected from six groups (12 participants, 1926 conversational turns). Results showed that SSRLBot can deliver detailed, theory-aligned evaluations, link specific behaviors to SSRL dimensions, and offer actionable recommendations for improving teamwork. The findings address a critical gap in medical education, advancing the application of LLM agents to enhance team-based decision-making and collaboration in high-stakes environments.

Keywords: Generative Artificial Intelligence · Large Language Models · Socially Shared Regulation · Medical Education · Conversational Agents

1 Introduction

Team actions–decisions and operations carried out collaboratively–are essential for achieving success. However, team action alone may not guarantee success without the joint efforts of team members in adopting appropriate strategies,

X. Huang and J. Gao—Equal contributions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 61–68, 2025.
https://doi.org/10.1007/978-3-031-99264-3_8

operations, and regulations [15]. This is especially true in high-stakes decision-making contexts, such as clinical reasoning, where a single decision can lead to significantly different outcomes [5]. The effectiveness of team decision-making is closely tied to the Socially Shared Regulation of Learning (SSRL, [13]). This framework emphasizes the collaborative regulation of learning processes as key to improving decision-making outcomes. In medical teamwork, team members' interactions through conversation play a pivotal role in their diagnostic outcome, as they could either direct to the correct or an opposite result [2]. Medical simulation has been pivotal for medical professionals to practice diagnostic skills and paves the way for understanding how teamwork works in this field [3,14].

In this study, we designed and developed a large language model (LLM)-based agent, SSRLBot. SSRLBot can evaluate students' performance in medical case diagnostics and conversations, judge the diagnosis results, and identify students' SSRL skills in their conversations. Grounded on the theory of SSRL, SSRLBot enables students to reflect on their diagnostic performance as a team and the key SSRL skills used to foster team diagnostic success in medical simulation. Additionally, we used six groups' diagnostic conversation data, including 12 participants and 1926 conversational turns, to evaluate this LLM-based agent.

2 Background and Related Work

2.1 SSRL in Team Work

Socially shared regulation is essential in teamwork. It involves team members' mutual efforts in regulating team performance through effective interactions. Socially shared regulated learning reflects learning strategies in educational contexts where two or more learners are involved. Studies in collaborative learning have found that groups that perform better show more SSRL strategies in teamwork [10]. However, the mechanism of SSRL is complex and hard to be captured, as dynamic interpersonal interactions occur where team members are influencing each other. SSRL guides interactions in teamwork from four dimensions, namely meta-cognitive, socio-cognitive, socio-motivational, and socio-emotional aspects. The meta-cognitive aspect of interactions (short in meta-cognitive interaction in the following content) refers to conversations about the team's cognitive progress [6,12]. Cognitive interaction includes conversations about the team's actual cognitive activities, such as planning, performing, and reflecting on the task and team performance [12]. Meta-cognitive interaction to teamwork is like an umbrella to an outdoor walking with a weather forecast; it prevents rains when necessary and can be folded when the sun comes out. Socio-emotional interaction, or social-emotional interaction, is team members' efforts to maintain cohesive and respectful social interaction through team dynamics [8]. Healthy and positive socio-emotional interaction enables team members to engage more effectively in cognitive interaction and leads to better team outcomes. Socio-motivational interaction is often discussed with socio-emotional interactions [8,11], as it is pivotal to motivating team spirits and triggers team members to stay tuned and engaged in teamwork.

SSRL research often conducts discourse analysis in computer-supported collaborative learning (CSCL) contexts, as it allows for the identification of learners' situational social, cognitive, emotional, and motivational states through transcriptions of their dialogues or written texts [7]. Integrating trace data from think-aloud protocols and CSCL logfiles represents a significant advancement in capturing learners' thought processes and actions, providing a more objective perspective on SSRL [1]. Most studies analyze verbal data through human coding, which, despite expert evaluation, remains labor-intensive and prone to bias. LLMs learn word usage patterns and apply them to natural language processing tasks [19], offering a more efficient alternative that often matches or surpasses human performance in content understanding, idea generation, and decision-making.

2.2 LLM Agent in Education

LLM agents have recently debuted in educational contexts. For example, PBChat is designed to identify and provide solutions to students' problem behaviors using teacher-parent conversations. LLMAgent-CK [21] is designed as a multi-agent framework to identify middle-school teachers' mathematical content knowledge learning goals from their responses. Other LLM-based agents are designed as virtual tutors or assistants in diverse subjects (e.g., Ruffle & Riley: Biology [16]; Jill Watsons, course readings [18]). Furthermore, other LLM-based agents have been designed for advancing educational research (e.g., Vizchat for visualizing multimodal learning analytics [20]). Despite previous applications, their potential for medical education remains unexplored. The SSRL theory-driven design could provide valuable insights into teamwork evaluation. Notably, SSRLBot is the first LLM-based agent designed to assess medical teams' SSRL interactions and their impact on diagnostic performance.

3 SSRLBot: System Architecture

Figure 1 shows the entire diagram from data collection to output evaluation, consisting of three key sections: 1) the theory-based development section, 2) the tuning section, and 3) the evaluation section. The architecture of SSRLBot integrates the GPT model (e.g., GPT-4), instructions, and SSRL rubrics. SSRLBot aims to improve the progress of data processing and optimize the quality of the data, reducing human-caused errors in data annotating and diagnosing. In addition, SSRLBot can provide actionable insights for researchers to improve their abilities in assessing and developing SSRL skills in educational, clinical, and collaborative learning environments.

3.1 Design

Since OpenAI launched the GPT store, GPT apps with different purposes have amplified the functionality of ChatGPT, making ChatGPT fully applicable to

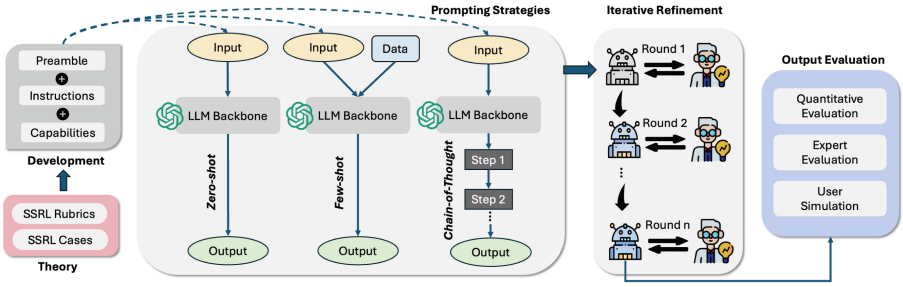


Fig. 1. The Workflow of the System Architecture for SSRLBot.

meet the needs of various users. GPT app configuration consists of four core components, including instructions, knowledge, capabilities, and actions. Instructions refer to development details of the GPT app configuration, such as the app's name, description, logo, instruction description, and conversation starters. Knowledge refers to the external knowledge files that can help improve response accuracy, fairness, and user experience [22]. Users can download these files if the code interpreter is enabled. Capabilities refer to web search, canvas, DALL.E Image Generation, and code interpreter and data analysis. Actions refer to make a third-party APIs to the GPT app. SSRLBot focuses on the instructions and capabilities of the GPT app framework. Furthermore, the SSRL framework is used to support specific functions under the instructions.

3.2 Functional Realization

The instruction description consists of a preamble and function descriptions. The preamble's setting helps SSRLBot clearly understand its role and continue playing this role in a scene for dialogue. We set the SSRLBot as a professional researcher with enough SSRL knowledge and health knowledge, be able to analyze and evaluate data. Moreover, we developed three core functions in SSRLBot: (1) identify the participant's interpersonal influence on each, (2) compare the SSRL skills, and (3) summarize and provide suggestions on participants. Figure 2 illustrates functions of SSRLBot. To enable SSRLBot to identify corresponding SSRL skills effectively and accurately, we developed an SSRL framework and integrated it into the system's preamble. This framework consists of three SSRL skill layers. The first layer contains five core SSRL skills: meta-cognitive interaction, socio-cognitive interaction, socio-motivational interaction, socio-emotional interaction, and task execution. Each core SSRL skill also contains the second and third SSRL skill layers. Definitions and examples were provided for each skill to help the system make more accurate judgments.

Although the SSRLBot has been set with three core functions to achieve the goal, the output is not guaranteed to be completely accurate. Tuning plays a key role in building the SSRLBot and enhancing dialogue competencies and accuracy. The tuning section in this study focuses primarily on two parts: instruction

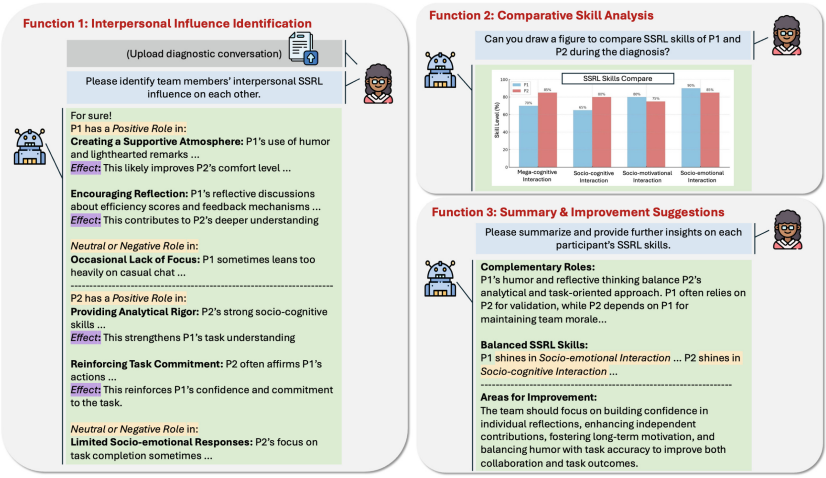


Fig. 2. The Demonstration of SSRLBot Functions.

and prompt. Instruction tuning helps provide feedback to system development and optimize instructions and preamble. Prompt tuning helps improve input quality. Previous studies [4,17,23] have demonstrated that using more appropriate prompts can improve output quality and mitigate hallucinations. This study conducted four rounds of iterative refinement. The first two rounds focused on functional completeness. The third round was mainly adjusted for output accuracy and quality. The fourth round focused on prompt improvement.

4 Illustrative Case Study

In this section, we evaluated SSRLBot using six groups of authentic conversational data, which were collected from 12 medical residents recruited from a university in Montreal, Canada.

Data Collection. Participants were paired up and required to discuss a medical case through the BioWorld system. BioWorld is a computer learning system that can provide a hospital simulation for students to discuss diseases using related knowledge [9]. The duration of each conversation was around 30 min. The number of conversational turn in each group was between 123 and 600. We conducted basic data cleaning on the collected data. For example, we added a speaker (e.g., P1 for participant 1, P2 for participant 2, R for researcher) and the corresponding SSRL skill on each conversational turn. In addition, we evaluated the diagnostic results of each group with the standard diagnostic answer to determine whether each group's results were correct.

Evaluation. Each group's conversational data was saved as a case in a single file. In this study, we mainly evaluated the diagnostic performance of SSRLBot. We requested SSRLBot summarize the conversation content for each case and evaluate the diagnostic results. The responses provided by SSRLBot primarily include three sections: a summary of the conversation, an evaluation of diagnostic accuracy, and a conclusion. This agent enhanced the precision of its outcomes by extracting the keywords and critical conversational content as evidence. Researchers compared the SSRLBot's results and the standard results and found that SSRLBot demonstrated 100% diagnostic accuracy.

Comparison. We mainly compare the performance of SSRLBot with other LLMs, including ChatGPT-3.5, Deepseek-R1, and Gemini-1.5. All selected LLMs demonstrated their ability to understand the conversation's goal, identify key information, and generate a general report based on the given content. However, ChatGPT 3.5 relied less on contextualized descriptions for each participant, limiting its usefulness for individuals aiming to improve their teamwork skills. DeepSeek-R1 provided a more thorough evaluation by blending positive and critical feedback on participants' SSRL skills and their impact on diagnostic performance. However, its theoretical alignment could be improved by explicitly linking behaviors to SSRL dimensions and offering more personalized feedback. Gemini-1.5 provided contextualized descriptions; for example, it acknowledged humor's role within the team and aligned with SSRL theory. However, its analysis lacked actionable feedback for human-centered reflection. In contrast, SSRLBot can follow the SSRL framework to better identify the specific SSRL skills from the conversation while highlighting specific examples. Additionally, it provided a comprehensive analysis by linking evidence from the conversation to each participant's SSRL skills, guiding researchers on which specific SSRL skill layers to adapt for improved team decision-making.

5 Conclusion and Future Work

This paper introduced SSRLBot, an LLM-based agent for medical education practice grounded in the SSRL framework. SSRLBot aims to facilitate customized feedback for learners through evaluating their conversations in the collaborative learning environment. This theory-driven knowledge-empowering design enhances LLM functions that support teamwork enhancement through customized feedback for individuals and groups. Our case study demonstrates that SSRLBot's responses are highly accurate. The limitation of this study is the limited number of cases for evaluation. Although each case included extensive conversational data, it is important to conduct more cases to further evaluate the accuracy and effectiveness of this agent. Future research should expand the sample size and evaluate other functions of SSRLBot. In addition, future research can enhance SSRLBot by expanding its application beyond dyadic medical diagnosis to diverse contexts, enabling a more comprehensive evaluation of its theory-driven functions. Additionally, while researchers assessed its contextual functions, incorporating participant feedback would provide a more robust

measure of its reliability and informativeness. Overall, the functionalities of this agent can be beneficial for medical education by helping assess learners' performance and SSRL skills and providing effective summarization and evaluation evidence for researchers.

References

1. Azevedo, R., et al.: Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. International Handbook of Metacognition and Learning Technologies, pp. 427–449 (2013)
2. Ball, J.R., Miller, B.T., Balogh, E.P.: Improving diagnosis in health care (2015)
3. Davis, J., et al.: Health care professional students' perceptions of teamwork and roles after an interprofessional critical care simulation. Dimens. Crit. Care Nurs. **40**(3), 174–185 (2021)
4. Diao, S., Wang, P., Lin, Y., Pan, R., Liu, X., Zhang, T.: Active prompting with chain-of-thought for large language models. arXiv preprint arXiv:2302.12246 (2023)
5. Djulbegovic, B., Guyatt, G.H.: Progress in evidence-based medicine: a quarter century on. The Lancet **390**(10092), 415–423 (2017)
6. Haataja, E., Dindar, M., Malmberg, J., Järvelä, S.: Individuals in a group: metacognitive and regulatory predictors of learning achievement in collaborative learning. Learn. Individ. Differ. **96**, 102146 (2022)
7. Hernández-Sellés, N., Muñoz-Carril, P.C., González-Sanmamed, M.: Computer-supported collaborative learning: an analysis of the relationship between interaction, emotional support and online collaborative tools. Comput. Educ. **138**, 1–12 (2019)
8. Järvelä, S., et al.: Enhancing socially shared regulation in collaborative learning groups: designing for CSCL regulation tools. Educ. Tech. Res. Dev. **63**, 125–142 (2015)
9. Lajoie, S.P., Guerrera, C., Munsie, S.D., Lavigne, N.C.: Constructing knowledge in the context of bioworld. Instr. Sci. **29**, 155–186 (2001)
10. Malmberg, J., Järvelä, S., Järvenoja, H., Panadero, E.: Promoting socially shared regulation of learning in CSCL: progress of socially shared regulation among high- and low-performing groups. Comput. Hum. Behav. **52**, 562–572 (2015)
11. Näykki, P., Järvelä, S., Kirschner, P.A., Järvenoja, H.: Socio-emotional conflict in collaborative learning-a process-oriented case study in a higher education context. Int. J. Educ. Res. **68**, 1–14 (2014)
12. Nguyen, A., Järvelä, S., Rosé, C., Järvenoja, H., Malmberg, J.: Examining socially shared regulation and shared physiological arousal events with multimodal learning analytics. Br. J. Edu. Technol. **54**(1), 293–312 (2023)
13. Panadero, E., Järvelä, S.: Socially shared regulation of learning: a review. Eur. Psychol. (2015)
14. Rosen, M.A., et al.: Promoting teamwork: an event-based approach to simulation-based teamwork training for emergency medicine residents. Acad. Emerg. Med. **15**(11), 1190–1198 (2008)
15. Salas, E., Sims, D.E., Burke, C.S.: Is there a "big five" in teamwork? Small Group Res. **36**(5), 555–599 (2005)

16. Schmucker, R., Xia, M., Azaria, A., Mitchell, T.: Ruffle&riley: insights from designing and evaluating a large language model-based conversational tutoring system. In: International Conference on Artificial Intelligence in Education, pp. 75–90. Springer (2024)
17. Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: evolution, applications, challenges, and opportunities. ACM Comput. Surv. **55**(13s), 1–40 (2023)
18. Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., Goel, A.K.: Jill watson: a virtual teaching assistant powered by ChatGPT. In: International Conference on Artificial Intelligence in Education, pp. 324–337. Springer (2024)
19. Thirunavukarasu, A.J., Ting, D., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.: Large language models in medicine. Nat. Med. **29**(8), 1930–1940 (2023)
20. Yan, L., et al.: VizChat: enhancing learning analytics dashboards with contextualised explanations using multimodal generative AI chatbots. In: International Conference on Artificial Intelligence in Education, pp. 180–193. Springer (2024)
21. Yang, K., et al.: Content knowledge identification with multi-agent large language models (LLMs). In: International Conference on Artificial Intelligence in Education, pp. 284–292. Springer (2024)
22. Zhang, Z., Zhang, L., Yuan, X., Zhang, A., Xu, M., Qian, F.: A first look at GPT apps: landscape and vulnerability. arXiv preprint arXiv:2402.15105 (2024)
23. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)

TPRS: AI-Assisted Research Topic Refinement for Distance Learners

Nghia Duong-Trung[1,2(✉)], Xia Wang[1], Rahul Rajkumar Bhoyar[1], Angelin Mary Jose[1], Silke Elisabeth Wrede[3], Lars van Rijn[3], Theresa Panse[3], Claudia de Witt[3], and Niels Pinkwart[1,4]

[1] German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91 C, 10559 Berlin, Germany
{nghia_trung.duong,xia.wang,rahul_rajkumar.bhoyar,angelin_mary.jose, niels.pinkwart}@dfki.de
[2] IU International University of Applied Sciences, Frankfurter Allee 73A, 10247 Berlin, Germany
[3] FernUniversität in Hagen, Universitätsstraße 27, 58097 Hagen, Germany
{silke.wrede,lars.vanrijn,theresa.panse,claudia.dewitt}@fernuni-hagen.de
[4] Humboldt-Universität zu Berlin, Unter den Linden 6, 10117 Berlin, Germany

Abstract. We present the Term Paper Recommendation System (TPRS), a hybrid AI-powered platform designed to scaffold the development of academic term papers in distance education. Tailored for students in a Bachelor of Arts program in Culture and Social Sciences, TPRS dynamically integrates large language models (LLMs), expert- and knowledge-based recommendation engines, and sentiment-driven routing to provide personalized formative feedback. A multi-shot prompting technique simulates high-fidelity tutoring interactions, trained on real supervision logs. Our system prioritizes transparency, student autonomy, and pedagogical alignment by combining structured validation, explainable recommendations, and relevance-based literature suggestions. A pilot deployment involving 18 students showed statistically significant improvements in submission quality (Hedge's $g = 0.44$, $p < .05$) and positive user reception across accuracy, usability, and trust dimensions, evaluated using the CRS-Que framework. This work contributes a modular, pedagogically informed approach to AI-supported academic writing, offering a promising direction for scalable, inquiry-driven support systems in higher education.

Keywords: Intelligent Tutoring · Generative AI · Distance Learning · Topic Refinement · CRS-Que

1 Introduction

Artificial Intelligence (AI) technologies have become integral to advancing instructional efficiency and addressing pedagogical challenges in higher education [1]. Applications of AI span a range of domains, including personalized

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 69–76, 2025.
https://doi.org/10.1007/978-3-031-99264-3_9

learning [6], intelligent tutoring [2], and automated feedback [13]. Pérez-Ortiz *et al.* conceptualize AI-based personal learning companions as tools capable of delivering adaptive, personalized support for learners [12]. A growing body of systematic reviews highlights the increasing adoption of AI technologies in higher education, particularly in promoting self-regulated learning and academic assistance [3,11]. The UNESCO AI Competency Frameworks for Teachers and Students emphasize the urgency of cultivating AI literacy, ethical awareness, and technical skills among educators and learners. The teacher framework focuses on AI pedagogy, ethics, and continuous professional development to ensure meaningful classroom integration [8], while the student framework centers on ethical responsibility, AI literacy, and foundational system design [9]. These initiatives collectively drive research into AI-driven educational systems, advocating for personalized learning, ethical AI deployment, and inclusive, sustainable educational transformation.

This work introduces the development prototype of TPRS, an AI-assisted tool designed to support German-speaking students in a Bachelor of Arts distance learning program in Culture and Social Sciences. The system aims to provide inspiring, personalized feedback on initial ideas for research topics and questions, which traditionally involve intensive one-on-one interactions between students and instructors. The existing workflow requires students to independently propose research topics and questions, which they submit via online platforms, such as Moodle, for iterative feedback from instructors. This process is labor-intensive and prone to delays. From a student's perspective, distilling complex academic materials into focused research ideas can be cognitively overwhelming, while long waiting times for feedback can fragment their workflow and reduce motivation. For instructors, the challenge lies in maintaining the depth and consistency of feedback while handling the administrative workload for large cohorts.

TPRS addresses this challenge through AI-powered knowledge scaffolding [4,14], offering structured yet flexible support that preserves student autonomy. Rather than delivering fully automated outputs, the system guides students in refining their research focus while encouraging critical thinking [7,10]. By combining LLMs with expert- and knowledge-based recommendation strategies, TPRS analyzes student input, provides instant formative feedback, and suggests relevant academic literature to support topic development. Unlike purely generative AI tools, TPRS is designed to balance automation with pedagogical alignment. It acts as a cognitive scaffold, not a replacement for academic reasoning. In a pilot deployment, we applied the CRS-Que framework [5] to evaluate system usability, accuracy, transparency, and user trust. The results indicate that TPRS improves student engagement, accelerates the feedback cycle, and reduces instructor workload, while also raising important questions about explainability and responsible AI use. This study contributes to prototyping the development of AI-driven educational support tools by demonstrating how hybrid recommendation systems can enhance academic self-regulation and support research topic refinement in distance learning environments.

2 TPRS System Design and Development

2.1 Challenges in the Research-Oriented Term Paper Process

In the final semester of a Bachelor of Arts program in Culture and Social Sciences, students are required to write a term paper that synthesizes knowledge from prior coursework, particularly in Media Education and Media Pedagogy. This research-oriented semester is structured into three phases, yet each presents specific challenges for both students and instructors. During the Term Paper Preparation phase, students are expected to independently reflect on course materials, identify thematic interests, and formulate a coherent research topic along with corresponding research questions. This step is cognitively demanding: the process of narrowing broad content into a specific, academically viable focus often leads to overload and uncertainty. Many students struggle to define clear, well-scoped questions without substantial guidance, resulting in frustration and delays. The Student-Teacher Interaction phase typically unfolds via a Moodle discussion forum, where students submit their initial ideas for feedback. However, feedback is rarely immediate. These delays can interrupt students' momentum and diminish motivation. Moreover, the varying quality of submissions often necessitates multiple rounds of revision, turning the feedback loop into a prolonged and sometimes demotivating experience. Once a topic is approved, students proceed to the Term Paper Writing phase. Yet, time lost in earlier stages often compresses the actual writing window, which can negatively impact the final paper's quality. Instructors, meanwhile, face their own set of challenges. Providing individualized feedback through Moodle for large cohorts is time-consuming and difficult to scale. Maintaining consistency and depth of feedback across students with diverse skill levels and research interests is especially demanding. Some students need extensive support, while others require minimal input, leading to an uneven distribution of effort. Additionally, tracking multiple rounds of revisions creates an administrative burden that detracts from more pedagogically meaningful engagement.

2.2 System Architecture and Workflow

The TPRS workflow begins when a student submits an initial term paper proposal consisting of a research topic (RT) and one or more research questions (RQ_i, where $i \geq 0$). A sentiment analysis (SA) module evaluates the student's confidence in their submission. Based on this assessment, the system dynamically selects one of two recommender engines: the knowledge-based engine (KE) or the expert-based engine (EE). Confident submissions are routed to the KE, which recommends literature to deepen academic exploration. In contrast, submissions with low confidence activate the EE, which offers more foundational guidance, such as refining the topic or restructuring the questions. Students may also override the system's decision and manually select a preferred engine or language model (LLM), including ChatGPT-4o, Mixtral-8x7B-Instruct-v0.1, and Llama-3-8B. Following this, the system conducts structured validation using

a set of evaluation modules (Eva_X) that assess the coherence and feasibility of the research topic (Eva_{RT}), each individual question (Eva_{RQ_i}), and topic-question pairs ($\text{Eva}_{\text{RT,RQ}_i}$). These validators are currently implemented using structured prompting techniques with ChatGPT-4o. Based on the validation results, the system generates positive or negative feedback accompanied by brief explanations to ensure transparency and trust. Two core recommendation scenarios are currently supported: (1) if the research topic is invalid but the questions are valid, the expert-based engine suggests a new topic aligned with the questions; (2) if both topic and questions are valid, the knowledge-based engine recommends up to three relevant academic sources to support further inquiry. Future iterations aim to expand this logic and generate more granular feedback across a broader range of submission patterns. To enhance recommendation quality, TPRS employs a multi-shot prompting strategy informed by historical student-teacher interactions. The system has been trained on a dataset of approximately 70 students from the same BA program, each with an average of 13âĂŞ15 exchanges. These interactions, collected in 2023 and the Winter 2023/2024 semester, are embedded into prompt templates to simulate high-fidelity tutoring. The resulting prompts are processed by a selected LLM, which generates structured, actionable feedback to help students refine their ideas. For literature recommendations, the system applies a cross-lingual semantic similarity algorithm. It generates a multilingual embedding of the student's topic using a model such as Multilingual BERT, then computes cosine similarity between this embedding and entries in a local academic repository. The most relevant references are retrieved and presented, ensuring that literature suggestions align closely with the student's research focus.

3 Pilot Evaluation

We conducted a pilot evaluation of TPRS within a Bachelor of Arts distance-learning program in Culture and Social Sciences at FernUniversität in Hagen. The study took place from 01.11.2024 to 20.12.2024 and included 18 student participants spanning six age groups. The majority identified as female (77.78%), with male (16.67%) and non-binary (5.55%) participants also represented. Most students (66.67%) were enrolled part-time, while the remainder studied full-time. Despite variations in demographic background and AI familiarity, participants had comparable academic progress, with most having completed 8 to 9 course modules.

To evaluate the system's usability and pedagogical impact, we applied the CRS-Que framework [5], which assesses conversational recommender systems along four dimensions: Perceived Qualities, User Beliefs, User Attitudes, and Behavioral Intentions. *Perceived Qualities* received good ratings across three criteria: accuracy, novelty, and explainability. Students found the recommendations relevant and helpful, with many highlighting their novelty and unexpected insights as a strength. The system's ability to explain its recommendations clearly contributed to user trust and transparency. *User Beliefs* were

measured in terms of perceived ease of use, usefulness, and control. Participants reported that TPRS was intuitive and effective for academic planning tasks such as topic development and literature search. They appreciated having the option to choose among different engines and LLMs, reinforcing a sense of control. *User Attitudes*, including trust and satisfaction, were also positive. Students expressed confidence in the system's guidance and reported that it helped clarify their research direction. Satisfaction scores indicated a fair level of approval, particularly for feedback responsiveness. Finally, in the dimension of *Behavioral Intentions*, most students indicated a willingness to reuse TPRS for future assignments and recommend it to peers. Besides the CRS-Que framework, the authors added *Judgment on Quality of Work* dimension by asking the students: How would you rate the quality of your work on the three main tasks of finding a topic for your homework in the module? Please rate individually in the case study, theoretical references, selected topic, and central question.

As hypothesized, analysis of submission quality demonstrated that supporting students in their term paper planning process has the potential to increase submission quality. This increase in quality may have effects on the intensity of the following tutoring process and may yield efficiency gains in either time spent on or the depth of tutoring, benefiting both students and teachers in the long run. To assess academic impact, domain experts evaluated students' initial and final submissions using a structured rubric. The rubric included four academic criteria: use case description (3 points), theoretical background (3 points), research topic formulation (2 points), and quality of research questions (3 points), on a total 11-point scale. Two independent raters ensured interrater reliability, achieving substantial agreement (Cohen's $\kappa = 0.675$, $p < .001$). Discrepancies were resolved through consensus. A directional paired t-test revealed a significant improvement in submission quality from initial ($M_{t1} = 6.79$, $SD_{t1} = 2.07$) to final ($M_{t2} = 7.53$, $SD_{t2} = 1.74$) drafts, with a small to moderate effect size (Hedge's $g = 0.44$, $p < .05$). A Shapiro-Wilk test confirmed that the initial scores deviated from normality, but further analysis retained all data points. Importantly, demographic covariates (age, gender, study mode, academic progress) showed no significant correlation with the observed improvement. These findings indicate that TPRS positively influenced the quality of students' academic submissions while offering a scalable model for reducing instructor workload and supporting inquiry-based learning.

4 Technical Report

The TPRS has been developed as a fully integrated system comprising both frontend and backend components. The system facilitates seamless interactions between students and AI-powered feedback mechanisms, replacing the conventional "to be reviewed by teachers" and "teachers post feedback" blocks in the workflow with "to be reviewed by AI" and "AI's instance feedback," respectively. The TPRS system was designed and implemented from scratch to ensure scalability, modularity, and compliance with educational and data protection standards.

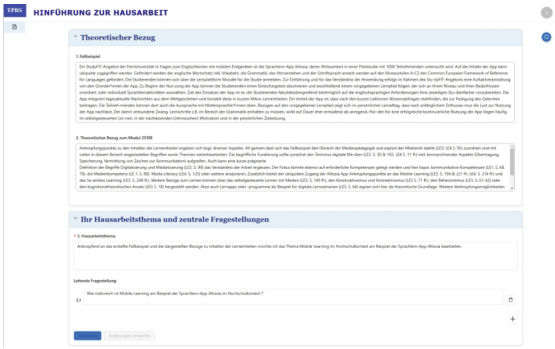


Fig. 1. The TPRS interface: Students' term paper ideas, reflection on learning resources, and several research questions.

During the pilot study, students interacted with TPRS using three selectable LLMs. The most frequently chosen model was Mixtral-8x7B-Instruct-v0.1 (54 interactions), followed by ChatGPT-4o (41 interactions) and Llama-3-8B (37 interactions). On average, each interaction involved approximately 2,606 input tokens and 472 output tokens. Mixtral produced the most consistent outputs, typically around 512 tokens, while ChatGPT-4o generated more variable responses, with a peak of 648 tokens. Llama-3-8B generally produced shorter outputs, reflecting a more conservative response style. The total cost of deploying TPRS during the pilot was $1004.59, primarily incurred through AWS services. The largest expenses stemmed from OpenSearch, SageMaker, and EC2 instances, which together accounted for over 80% of the total. OpenSearch alone made up nearly 40%. Other services used–such as Lambda, CloudWatch, and Secrets Manager–were part of AWS's free tier. The cost analysis confirms that the system is feasible for medium-scale deployment in educational settings.

The TPRS frontend provides students with an intuitive interface to submit proposals, select models, and view feedback history. Built using HTML, CSS, JavaScript, and Node.js, the interface adheres to accessibility standards and supports personalized feedback tracking. Package management is handled via YARN, and user data is stored using MongoDB. The backend is implemented with FastAPI for efficient RESTful API handling, and LangChain is used to manage interactions across multiple LLMs. ChatGPT-4o is accessed via OpenAI's API, while open-source models are integrated through Hugging Face. A modular architecture allows easy integration of new models or APIs in future iterations. Continuous integration and deployment (CI/CD) pipelines are in place to support iterative improvements. The system is hosted on AWS infrastructure within the EU to ensure compliance with GDPR and other data protection regulations. Several screenshots of TPRS are illustrated in Figs. 1 and 2. TPRS also integrates with the university library system, enabling retrieval of academic references aligned with students' topics. This connection ensures that literature recommendations are pedagogically grounded and domain-specific.

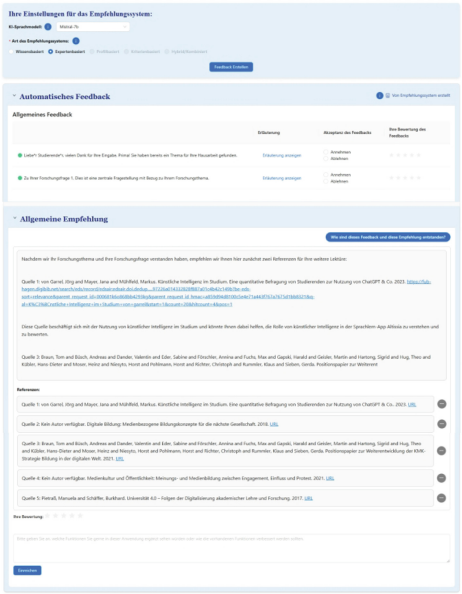


Fig. 2. The TPRS interface: Automatic feedback and recommendation.

5 Conclusion

This paper presented TPRS, an AI-assisted system designed to scaffold the development of term paper topics and research questions in distance learning environments. Combining LLMs with knowledge-based and expert-driven recommendation strategies, TPRS adapts its feedback based on students' confidence levels and leverages historical supervision data to simulate high-fidelity guidance. Structured validation modules and semantic similarity algorithms ensure that feedback and literature recommendations are contextually relevant and academically sound. Through a pilot deployment with students in a Bachelor of Arts program, TPRS demonstrated its ability to improve submission quality while reducing instructor workload. Evaluation using the CRS-Que framework confirmed high levels of perceived accuracy, usability, and trust. Technically, the system integrates modern web technologies, modular LLM orchestration via LangChain, and scalable cloud deployment on GDPR-compliant infrastructure.

Acknowledgments. The authors kindly appreciate the support of CATALPA, Center of Advanced Technology for Assisted Learning and Predictive Analytics, FernUniversität in Hagen, through the project "AI.EDU Research Lab 2.0".

References

1. Al-Zahrani, A.M., Alasmari, T.M.: Exploring the impact of artificial intelligence on higher education: the dynamics of ethical, social, and educational implications. Humanit. Soc. Sci. Commun. **11**(1), 1–12 (2024)
2. Bailón, A.B., Contreras, W.F., Molina-Solana, M.: Intelligent tutoring system, based on video e-learning, for teaching artificial intelligence. In: Pérez, J.B., et al. (eds.) Trends in Practical Applications of Agents, Multi-Agent Systems and Sustainability - The PAAMS Collection, 13th International Conference, PAAMS 2015, Salamanca, Spain, June 3–4, 2015, Special Sessions. Advances in Intelligent Systems and Computing, vol. 372, pp. 215–224. Springer (2015)
3. Castillo-Martínez, I.M., Flores-Bueno, D., Gómez-Puente, S.M., Vite-León, V.O.: AI in higher education: a systematic literature review. Front. Educ. **9**, 1391485 (2024)
4. Ertugruloglu, E., Mearns, T., Admiraal, W.: Scaffolding what, why and how? a critical thematic review study of descriptions, goals, and means of language scaffolding in bilingual education contexts. Educ. Res. Rev. 100550 (2023)
5. Jin, Y., Chen, L., Cai, W., Zhao, X.: CRS-que: a user-centric evaluation framework for conversational recommender systems. ACM Trans. Recommender Syst. **2**(1), 1–34 (2024)
6. Kaswan, K.S., Dhatterwal, J.S., Ojha, R.P.: AI in personalized learning. In: Advances in Technological Innovations in Higher Education, pp. 103–117. CRC Press (2024)
7. Kim, J., Lee, H., Cho, Y.H.: Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education. Educ. Inf. Technol. **27**(5), 6069–6104 (2022)
8. Miao, F., Cukurova, M.: AI competency framework for teachers (2024)
9. Miao, F., Shiohira, K.: AI competency framework for students (2024)
10. Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., Paquette, L.: Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. J. Comput. Assist. Learn. **39**(2), 351–368 (2023)
11. Ouyang, F., Zheng, L., Jiao, P.: Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. Educ. Inf. Technol. **27**(6), 7893–7925 (2022)
12. Perez-Ortiz, M., Novak, E., Bulathwela, S., Shawe-Taylor, J.: An AI-based learning companion promoting lifelong learning opportunities for all. arXiv preprint arXiv:2112.01242 (2021)
13. Shi, H., Aryadoust, V.: A systematic review of AI-based automated written feedback research. ReCALL 1–23 (2024)
14. Zeng, J., Zhang, P., Zhou, J., Shang, J., Black, J.B.: The impact of embodied scaffolding sequences on stem conceptual learning. Educ. Technol. Res. Dev. 1–26 (2024)

Facilitating Teacher Coaching: Generating Insightful Comments on Classroom Instruction Through Multi-agent Reflection

Jinglei Yu[1] and Yu Lu[1,2(✉)]

[1] School of Educational Technology, Faculty of Education, Beijing Normal University, Beijing 100875, China
luyu@bnu.edu.cn

[2] Advanced Innovation Center for Future Education, Beijing Normal University, Beijing 100875, China

Abstract. Video-based classroom analysis offers great opportunities for teachers to reflect on their own or others' performance for professional development. However, the substantial time and resource demands of expert or peer insights restrict teachers' access to consistent instructional support. Large language models (LLMs) have the potential to generate content-specific comments based on the classroom transcript, but their insights often lack novelty and expert professional vision. To address this, we propose an LLM-based comment generation mechanism through multi-agent reflection, which simulates expert instruction by integrating professional vision and rubric-based feedback. Experimental results demonstrate that this mechanism can effectively generate insightful comments with deeper analyses of student thinking and more specific suggestions for teaching improvement, facilitated through iterative revisions between LLM agents in the reflection process. The study presents an effective method for generating insightful feedback for teachers by simulating expert guidance, which is expected to enhance the practical value of generative AI for teacher coaching.

Keywords: Multi-Agent Reflection · Large Language Model · Teacher Professional Vision · Teacher Learning · Classroom Analysis · Generative AI

1 Introduction

In teacher education, video-based classroom analysis offers teachers a valuable approach to capturing classroom events and reflecting on instructional performance. Engaging in video clubs for peer discussions [1] or receiving expert feedback [2] supports teacher change and enhances teacher noticing. However, the substantial time required from peers and experts limits consistent instructional support.

With the rapid advancement of generative AI, large language models (LLMs) show significant potential for generating human-like comments. Recent studies have explored LLM applications in classroom analysis using zero-shot prompting. For instance, Wang and Demszky [3] utilized ChatGPT to detect key moments and overlooked aspects

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 77–84, 2025.
https://doi.org/10.1007/978-3-031-99264-3_10

in classroom instruction and offer suggestions for improvement based on classroom transcript. Yu et al. [4] proposed a framework that analyzes student behavior through human pose estimation in videos and generates textual insights using LLMs. Hou et al. [5] employed ChatGPT to score observation protocol rubrics and provide logical justifications for scores based on classroom transcript.

These studies demonstrate the feasibility of integrating LLMs into classroom analysis. However, the quality of LLM-generated comments still requires further improvement to be effectively utilized in teacher coaching. Without an effective approach to embedding expert instruction, LLM-generated feedback risks lacking novelty and depth [3]. Therefore, this study aims to integrate expert instructions into LLM-based comment generation, simulating expert guidance for LLMs to produce more insightful comment on classroom instruction for teacher coaching.

2 LLM-Based Comment Generation for Teacher Coaching

We propose an LLM-based multi-agent reflection mechanism to generate comments on classroom instruction for teacher coaching (see Fig. 1). Compared to the conventional practice of human discussion and feedback, our approach employs LLM-based agents to generate comments from classroom transcript, reducing the heavy reliance on human teachers. To augment LLMs with professional knowledge for insightful comment generation, expert instructions for the LLMs are embedded in the system messages of each agent, namely the comment-writing agent and the comment-review agent. We will provide further details about the proposed mechanism in the following sections.

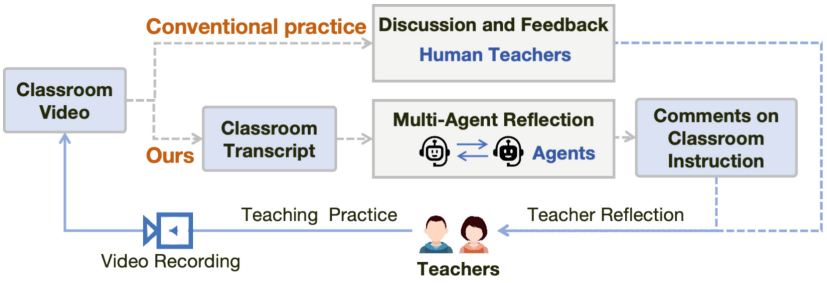


Fig. 1. The illustration of our LLM-based comment generation mechanism and conventional practice for teacher coaching.

2.1 Expert Instruction for Comment-Writing Agent

The comment-writing agent is prompted to act as a professional in-service teacher, writing comment on classroom instruction and collaborating with the comment-review agent for refinement. To incorporate expert instruction, the comment-writing agent formulates comments based on three specialized roles of awareness (see Fig. 2):

Situation Awareness. Describing critical events is fundamental to classroom analysis for teacher noticing [6]. Teachers are encouraged to focus on specific aspects with evidence, fostering reflective thinking [7]. Therefore, we adopted a mentor-used prompt to guide teachers in detailing critical events [8] for prompting the LLMs.

Student Awareness. Attending to and interpreting students' thinking is essential for teacher noticing [9]. Compared to novices, expert teachers check on all students more frequently during classroom analysis [10]. Thus, to guide LLMs in adopting a comprehensive focus on students, we integrated the SOLO taxonomy [11], enabling a multi-level assessment of student learning outcomes.

Knowledge Awareness. Applying knowledge to analyze and reason about critical events is a core aspect of noticing for teachers [6]. This knowledge-driven analysis is also one of distinction between expert and novice teachers [12]. Thus, we prompt LLMs to identify teachers' practical knowledge [13] and evaluate it using its pre-trained knowledge instead of a specific database, allowing for flexible adaptation to the context.

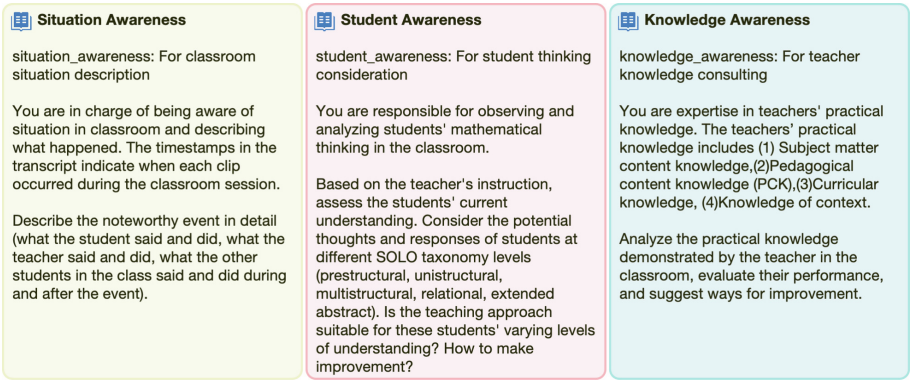


Fig. 2. Three key components of the comment-writing agent's system message: situation awareness (left), student awareness (middle), and knowledge awareness (right).

2.2 Rubric Scaffolding for Comment-Review Agent

The comment-review agent is responsible for evaluating the generated comment and providing suggestions for improvement. Evaluation rubrics can effectively support peer feedback training by helping teachers focus on key content areas [14]. Therefore, we adopt a rubric for assessing teachers' ability to analyze classroom instruction (see Table 1), enabling the review agent to assign scores (0–10) and suggest improvements.

Table 1. Rubric for assessing teachers' ability to analyze classroom instruction [15].

Category	Content
Rubric 1	Attend to the mathematics content at the center of the interaction
Rubric 2	Elaborate on students' mathematical thinking and learning as made evident in the video clips
Rubric 3	Propose improvements
Rubric 4	Elaborate and interpret the interaction observed in the video beyond a mere description of what was observed

2.3 Multi-agent Reflection via AutoGen Framework

The multi-agent reflection mechanism, shown in Fig. 3, is adapted from the AutoGen multi-agent design pattern [16]: reflection for code writing. Messages are dynamically published and optionally handled by two agents based on message types. Specifically, the process begins with the comment-writing agent generating an initial comment based on the comment-writing task and constructing a corresponding comment-review task. The comment-review agent then evaluates the comment and publishes a review result. When the comment gained 9 or 10 scores on each rubric in the review result, the review agent can approve the comment. If the comment is not approved, the comment-writing agent must revise it based on the comment-review agent's feedback and reconstruct the review task. This iterative process continues until the reviewer's evaluation is "approved", finalizing the comment-writing result.

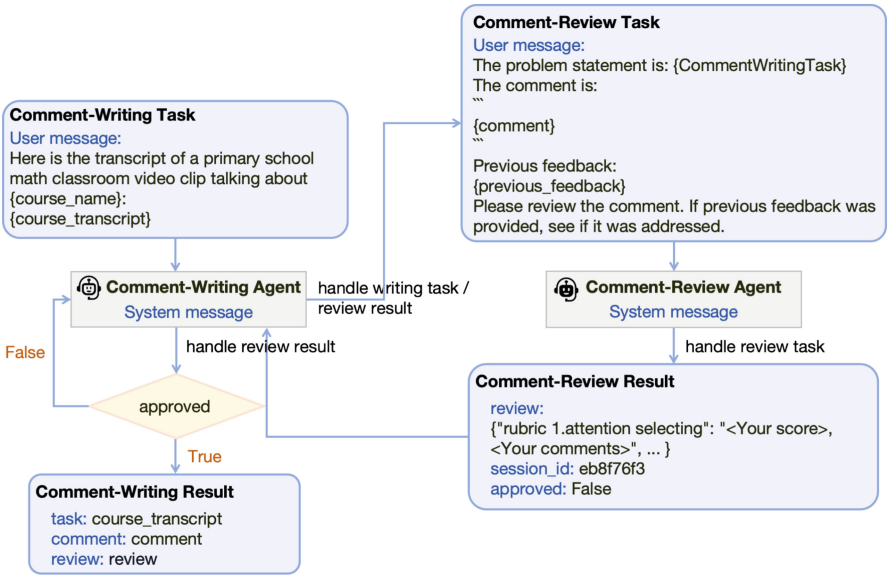


Fig. 3. The illustration of multi-agent reflection mechanism.

3 Experiments

Experiments were conducted from two perspectives. First, we compared the multi-agent reflection mechanism with single LLM generation to assess its effectiveness. Second, we analyzed multiple rounds of comments generated through the reflection mechanism to explore how it enhances comment generation.

3.1 Dataset

The dataset, including classroom transcript and timestamped human comments, were collected from previous teacher coaching projects and used solely for research. We sampled 5 elementary math classes, selecting two transcript segments per class, totaling 10 segments. The selected segments met three criteria: (a) enabling analysis of student thinking through teacher-student interactions, (b) containing multiple human comments, representing them noteworthy for teacher to attend to, and (c) presenting complete teaching events for contextual clarity. Each segment averages 1:56 min and includes comments from at least two in-service teachers or experts. For example, one segment is from the course "Parallel and Vertical", shown between "36:42 - 38:06" at the end of the 45-min class. During this segment, the teacher asks students to discuss real-life examples of vertical situations, selecting students individually to respond. Below are two human comments (translated into English) with their respective annotated timestamps:

- *"36:53": "Relate to life, discuss vertical situations in real life."*
- *"37:53": "Encourage students to observe perpendicular situations in life and view the real world through a mathematical lens."*

3.2 Metrics

Similarity to Transcripts. It calculates the semantic similarity between the generated comment and the classroom transcript using cosine similarity, which reflects how specifically the comment describes the teaching event.

Similarity to Human Comment. It calculates the mean semantic similarity between the generated comment and human comments using cosine similarity, which indicates whether the generated comment generally aligns with the intended human meaning.

Sentence Diversity. It calculates the mean value of pairwise semantic similarities between all sentences in the generated comment. Since lower similarity indicates higher diversity, the final result is obtained by subtracting the value from 1.

Rubric Content Distribution. It encodes each sentence in the generated comment with one or more rubrics. The coding framework aligns with the rubrics in Table 1.

3.3 Implementation Details

We used Deepseek-v3 [17] as the LLM for comment generation and M3e model [18] for Chinese text embedding. The fourth metric was initially coded with Deepseek-v3,

followed by human review and statistical analysis. The implementation of reflection mechanism was constrained to two rounds, meaning that if only one round was approved, the process would be retried. For comparison, single LLM generation was implemented using a single prompt, where the LLM was instructed to role-play based on the reflection mechanism description. The system messages for each role remained consistent with those in the reflection mechanism.

3.4 Results

Table 2 shows the results of the first three metrics for the reflection mechanism in round 1 and round 2, as well as single LLM generation. The round 2 comments show the longest length and achieve the highest mean similarity to transcripts, indicating that the content generated by the reflection mechanism is more specific and closely aligned with the original classroom content. In terms of similarity to human comments, the single LLM generation achieves the highest similarity, while the reflection mechanism maintains a consistent similarity. This suggests that, while the reflection mechanism stimulates more content, it still retains alignment with human evaluations. Finally, round 2 comments exhibit the highest sentence diversity, indicating that the mechanism fosters more varied content rather than simple repetition of original phrasing.

To assess improvements from round 1 to round 2, Table 3 shows enhancements in Rubric 2 and Rubric 3, particularly in Rubric 3. This indicates an increase in the depth of student thinking analysis and instructional improvement suggestions for teachers.

Table 2. The mean value results of the first three metrics.

Variables	Number of segments	Comment length	Similarity to transcripts	Similarity to human comment	Sentence diversity
Round1	10	303.6	0.835	0.767	0.174
Round2	10	**430.2**	**0.841**	0.767	**0.194**
Single	10	243.7	0.826	**0.768**	0.160

Table 3. The results of the fourth metric: rubric content distribution.

Variables	Number of segments	Mean counts per segment			
		Rubric 1	Rubric 2	Rubric 3	Rubric 4
Round1	10	2.5	2.5	2	1.8
Round2	10	2.5	**2.7**	**3.3**	1.8

To further assess comment quality, we analyzed examples by comparing sentence excerpts from both rounds to examine how the comments improved based on the

reviewer's suggestions. Regarding comments on the "Area of a polygon" segment, the round 1 review suggested the following for Rubric 3. "*Finally, suggesting specific real-life examples or open-ended questions could make the proposed improvements more actionable.*" Then in round 2, the revised comment (translated into English) offers more concrete and comprehensive improvement suggestions, as demonstrated below in bold:

- *Round 1: "Additionally, teachers can consider incorporating more real-life examples to enhance students' ability to apply geometric knowledge."*
- *Round 2: "Additionally, teachers can consider incorporating more real-life examples,* ***such as calculating the floor area of a room or designing the layout of a flower bed****, to enhance students' ability to apply geometric knowledge.* ***According to constructivist learning theory, teachers can also encourage students to collaboratively construct knowledge through cooperative learning. For example, group members can explain their transformation methods to one another, thereby deepening their understanding."***

4 Conclusion and Future Work

This study proposes an LLM-based multi-agent reflection mechanism that integrates expert instructions to generate insightful comments on classroom instruction, which aims to enhance the practical value of generative AI for teacher coaching. Specifically, the two agents collaborate to simulate expert comments by following expert instructions for both comment writing and review. During this process, the reflection mechanism is designed to facilitate their iterative refinement of the comments.

Our findings indicated that the reflection mechanism enhanced the depth of student thinking analysis and provided more specific and comprehensive improvement suggestions. This can be attributed to the higher level of engagement of LLM enforced by iterative revisions. The mechanism enables dynamic interaction, allowing reviewers to provide targeted feedback to writers, leading to more precise and in-depth evaluations.

For the future work, from a generation perspective, incorporating multimodal classroom data can further enhance the authenticity of classroom representation for analysis. From a validation perspective, we plan to interview teachers to evaluate the quality of generated comments. Furthermore, we aim to explore methods for integrating the proposed mechanism into coaching activities through human-AI collaboration, alleviating the burden on teachers while preserving their essential role.

Acknowledgments. This research is supported by the National Natural Science Foundation of China (No. 62477003).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Tripp, T.R., Rich, P.J.: The influence of video analysis on the process of teacher change. Teach. Teach. Educ. **28**(5), 728–739 (2012)

2. Kleinknecht, M., Gröschner, A.: Fostering preservice teachers' noticing with structured video feedback: results of an online-and video-based intervention study. Teach. Teach. Educ. **59**, 45–56 (2016)
3. Wang, R. E., Demszky, D.: Is chatgpt a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. arXiv preprint arXiv: 2306.03090 (2023)
4. Yu, Z., Xie, M., Gao, J., Liu, T., Fu, Y.: From raw video to pedagogical insights: a unified framework for student behavior analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 21, pp. 23241–23249. AAAI, Menlo Park (2024)
5. Hou, R., et al.: Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and ChatGPT. In: International Conference on Artificial Intelligence in Education, pp. 60–74. Springer, Cham (2024)
6. Van Es, E.A., Sherin, M.G.: Mathematics teachers' "learning to notice" in the context of a video club. Teach. Teach. Educ. **24**(2), 244–276 (2008)
7. Griffin, M.L.: Using critical incidents to promote and assess reflective thinking in preservice teachers. Reflective Pract. **4**(2), 207–220 (2003)
8. Rotem, S.H., Ayalon, M.: Changes in noticing multiple dimensions in classroom situations among pre-service mathematics teachers. Teach. Teach. Educ. **121**, 103932 (2023)
9. Barnhart, T., van Es, E.: Studying teacher noticing: examining the relationship among preservice science teachers' ability to attend, analyze and respond to student thinking. Teach. Teach. Educ. **45**, 83–93 (2015)
10. Kosel, C., Holzberger, D., Seidel, T.: Identifying expert and novice visual scanpath patterns and their relationship to assessing learning-relevant student characteristics. Front. Educ. **5**, 612175 (2021)
11. Blggs, J.B., Collis, K.F.: Evaluating the Quality of Learning: the SOLO Taxonomy. Academic Press, New York (1982)
12. Wolff, C.E., Jarodzka, H., van den Bogert, N., Boshuizen, H.P.: Teacher vision: expert and novice teachers' perception of problematic classroom management scenes. Instr. Sci. **44**, 243–265 (2016)
13. Witterholt, M., Goedhart, M., Suhre, C., van Streun, A.: The inter-connected model of professional growth as a means to assess the development of a mathematics teacher. Teach. Teach. Educ. **28**(5), 661–674 (2012)
14. Alqassab, M., Strijbos, J.W., Ufer, S.: Training peer-feedback skills on geometric construction tasks: role of domain knowledge and peer-feedback levels. Eur. J. Psychol. Educ. **33**, 11–30 (2018)
15. Santagata, R., Yeh, C.: The role of perception, interpretation, and decision making in the development of beginning teachers' competence. ZDM Math. Educ. **48**, 153–165 (2016)
16. AutoGen Muti-Agent Design Patterns: Reflection. https://microsoft.github.io/autogen/stable/user-guide/core-user-guide/design-patterns/reflection.html. Accessed 11 Feb 2025
17. Liu, A., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
18. Wang, Y., Sun, Q., He, S.: M3e: moka massive mixed embedding model, https://github.com/wangyuxinwhy/uniem. Accessed 11 Feb 2025

Modeling Socially Constructed Knowledge Using Multimodal Machine Learning: A Case Study in K-12 AI Literacy Education Classroom

Md Biplob Hosen[1], Lujie Karen Chen[1(✉)], Sancharee Hom Chowdhury[1], Erin Bosarge[2], Na Gong[2], Woei Hung[3], Becky Tomaso[4], and Shenghua Zha[2]

[1] University of Maryland Baltimore County, Baltimore, MD, USA
{mhosen1,lujiec,schowdh2}@umbc.edu
[2] University of South Alabama, Mobile, AL, USA
emb2021@jagmail.southalabama.edu, {nagong,shzha}@southalabama.edu
[3] University of North Dakota, Grand Forks, ND, USA
woei.hung@und.edu
[4] Baldwin Preparatory Academy, Loxley, AL, USA
btomaso@bcbe.org

Abstract. Artificial Intelligence (AI) literacy is gaining momentum in K-12 education, with active learning and small-group activities emerging as promising effective pedagogical strategies. However, teachers face challenges in monitoring and understanding small-group dynamics, which are crucial for enacting effective pedagogical tactics and providing timely support. Although it is becoming increasingly feasible to collect multimodal data (e.g., audio and video) in classrooms, translating these complex data streams into pedagogically relevant insights remains a challenge. The study is situated in an K-12 AI literacy project-based learning context and uses multimodal data streams collected from real classrooms, focusing on the process of social construction of knowledge emerging from small-group activities, including on-task versus off-task engagement and sharing versus beyond-sharing behaviors. We explored state-of-the-art multimodal machine learning models and demonstrated its advantage compared with unimodal approaches relying solely on features extracted from verbal messages. The results demonstrate the promise of multimodal models to support teaching analytics, even with a modest dataset collected in a noisy environment. We discuss the potential of this work to empower educators by leveraging complex multimodal data streams to support the process of promoting meaningful knowledge construction in AI literacy classrooms. This work contributes to the emerging body of research on the application of AI in real-world K-12 AI literacy settings and demonstrates how AI can serve not only as a subject of instruction but also as a tool to advance AI literacy education in an ecologically valid educational context.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 85–92, 2025.
https://doi.org/10.1007/978-3-031-99264-3_11

Keywords: AI literacy · Social construction of knowledge · Multimodal analytics · Features extraction · Hybrid fusion · Machine learning

1 Introduction

AI education for K-12 students has grown exponentially in recent years. Existing studies focused on AI literacy curriculum design, instrument validation, and pedagogical approaches [6,18]. Among the three foci, examination on pedagogical approaches serves as the foundation for the other two and is also the most extensively studied. Individual instructional practices were deployed to identify optimal pedagogy on students' learning of specific AI concepts. For example, Lu and Fan [16] used project-based learning to teach high-school students supervised learning concepts to solve weather prediction problems. Dai et al. [8] used an embodied-analogical-disruptive approach to teach elementary students basic concepts of sensors and machine perception as well as machine learning.

Among the studies of pedagogical approaches, a majority of the results were generated from the analysis of knowledge tests, perception surveys, interviews, or artifact analysis conducted at the beginning and/or end of the AI education interventions [17]. While the results of comparison explained the effectiveness of interventions, they failed to capture the nuance of student development in the AI learning process. To gain a deeper understanding of how students engage in AI learning, it is crucial to analyze students' collaborative and individual learning processes through which knowledge is constructed and refined. This study used and compared multimodal machine learning (ML) methods to analyze students' social construction of knowledge in an AI literacy project-based learning lesson in a high-school class. Through comparative analysis of multimodal features and leave-one-video-out cross-validation, we demonstrated: (1) the significant advantage of multimodal analysis over unimodal approaches for capturing SCK processes, and (2) the real-world applicability of our method, with performance metrics approximating authentic classroom conditions.

2 Related Work

2.1 Social Construction of Knowledge

Social Construction of Knowledge (SCK) is an important concept in social psychology [12]. It refers to the idea that knowledge is constructed on social exchange among people in communities. Studies of SCK have built upon Gergen's [11] definition that knowledge was "represented in linguistic propositions" (p. 270), thereby focusing on examining learners' verbal behaviors, such as discussions in online forums or social media messages [14]. Among them, Gunawardena's [13] interaction analysis model was widely used in examining SCK. The model described five phases when SCK was developed, including information sharing,

identifying dissonance, negotiation, testing and modifying the proposed synthesis, and application of newly constructed meaning. Nguyen and Diederich [19] adopted this model and analyzed comments under Tiktok educational videos. They found that the highest number of comments had non-SCK purposes, like off-task socialization, followed by comments on information sharing. Videos with informational content drew more knowledge sharing and negotiation comments than other videos. Ye and Pennisi [23] modified Gunawardena's [13] model into a cyclic model, highlighting the continuous feature of SCK after students acquired and shared new knowledge. In addition, their results endorsed Nguyen and Diederich's [19] results that students' SCK mainly occurred through information sharing by reading and commenting on others' posts. While the studies of verbal interactions of SCK capture valuable insights on students' learning progress and engagement, the limitations of relying solely on one data source are evident. Researchers have identified concerns such as bias towards extroverted students [20]. In addition, annotations of verbal interactions were time and resource intensive. Therefore, it is essential to incorporate multiple data sources, such as non-verbal cues, to gain a more comprehensive understanding of students' SCK.

2.2 Multimodal Learning Analytics

Multimodal learning analytics (MMLA) has advanced significantly in both controlled and classroom settings. In controlled environments, studies demonstrate multimodal fusion's effectiveness: combined verbal-prosodic features improve problem-solving detection [4], and facial/gaze data enhances satisfaction prediction in game-based learning [1]. Zhao et al.'s METS framework [24] further shows how spatial-temporal data can model team dynamics. In real-world classrooms, researchers have developed practical applications including engagement classification systems [25] and team dynamics visualizations [9]. While these tools demonstrate MMLA's potential, challenges remain in adapting methods to noisy classroom environments. Our work addresses this gap through rigorous validation approaches suitable for authentic educational settings.

3 Method

3.1 Data Collection and Annotation

This study examined 16 high school students across five groups in an AI literacy project-based learning lesson at a public high school in the Southern U.S. All participants were informed in advance about the video recording and provided written consent. Prior to the lesson, students had learned basic AI concepts, such as supervised learning and neural networks. They also collected drawings from children at kindergarten and elementary schools. During the lesson, students were instructed to use Teachable Machine [5], an AI learning tool, to build a machine learning model that distinguished the drawings based on five stages of artistic development [15].

Three researchers conducted the manual annotation or qualitative coding based on the literature review of SCK. The coding focused on two coding categories: on/off-task interactions and knowledge-sharing/beyond-sharing. If a student's conversation did not focus on the project, it was annotated off-task conversation. Otherwise, it was coded as an on-task conversation. In addition, on-task conversations between students were annotated as knowledge sharing if students exchanged their understanding or knowledge of the involved concepts. Other on-task conversations were labeled as beyond sharing as they focused on SCK levels higher than knowledge sharing, such as negotiation. Individual coding and group meetings were conducted iteratively until 100% agreement (interrater reliability) was reached. The unit of annotation was the sentence.

3.2 Features Extraction and Fusion

Our multimodal feature extraction comprised: (1) textual analysis using LIWC-22 [3] (110 dimensions spanning emotional tone, cognitive complexity, and social interaction) and DistilBERT [22] (768-dimensional embeddings); (2) audio processing with openSMILE [10] (88 prosodic features) and Wav2Vec 2.0 [2] (32-dimensional embeddings via mean pooling); and (3) visual encoding via the CLIP image encoder [21] (510-dimensional embeddings from 30 FPS keyframes), all aggregated into unified representations for knowledge construction analysis.

Feature fusion can be implemented at various stages: early fusion combines features at the input level, late fusion merges decisions from separate models, and hybrid fusion incorporates both strategies [7]. In this work, we adopted a hybrid fusion approach that integrates both feature-level and decision-level fusion. At the feature level, we combined the best-performing verbal and audio features selected by their area under the curve (AUC) scores. For decision-level fusion, we implemented a meta-model to combine predictions from the feature-fused classifier with those from the visual modality classifier.

3.3 Model Training and Evaluation

We evaluated four classifiers (Logistic Regression, Support Vector Machine, Random Forest, and Multilayer Perceptron) using a stratified 70-30 train-test split. Hyperparameter tuning is performed via Grid Search to identify the optimal configuration. Models were optimized via 5-fold cross-validation with GridSearch for parameter tuning. This approach maintains the class distribution across folds to address class imbalances and avoid skewed results. Performance was assessed by AUC scores for two tasks: "on-task" versus "off-task" behaviors and "sharing" versus "beyond-sharing" tasks.

4 Result

4.1 On-Task vs Off-Task Model Performance

The results shown in Table 1 provide key insight into features performance and classifiers effectiveness for on-task vs. off-task classification. BERT-based verbal

features deliver the best performance for textual data, achieving an AUC of 0.744 in Logistic Regression. For audio, prosodic features lead, with Logistic Regression achieving an AUC of 0.660. Among unimodal features, CLIP visual embeddings outperform others. However, the best overall performance is achieved with hybrid fusion, combining BERT, prosodic, and CLIP features. Random Forest, trained on these hybrid multimodal features, yields the best result with an AUC of 0.875, outperforming all other models and feature combinations.

Table 1. AUC-scores for on-task vs off-task for different feature sets.

Classifier	Verbal Features		Audio Features		Visual Features	Fusion
	LIWC	BERT	Prosodic	Wav2Vec	CLIP	Hybrid
Logistic Regression	0.713	0.744	0.660	0.487	0.827	0.863
Support Vector Machine	0.709	0.721	0.655	0.514	0.828	0.872
Random Forest	0.724	0.742	0.632	0.489	0.847	0.875
Multilayer Perceptron	0.697	0.723	0.589	0.545	0.828	0.862

Table 2. AUC-scores for sharing vs beyond-sharing for different feature sets.

Classifier	Verbal Features		Audio Features		Visual Features	Fusion
	LIWC	BERT	Prosodic	Wav2Vec	CLIP	Hybrid
Logistic Regression	0.714	0.677	0.554	0.560	0.868	0.924
Support Vector Machine	0.674	0.759	0.484	0.520	0.921	0.943
Random Forest	0.661	0.685	0.486	0.600	0.871	0.905
Multilayer Perceptron	0.651	0.668	0.482	0.558	0.902	0.916

4.2 Sharing vs Beyond-Sharing Model Performance

For the sharing vs. beyond-sharing classification (Table 2), BERT-based verbal features provide the best performance, with an AUC of 0.759 using SVM, outperforming LIWC (0.714 with Logistic Regression). In audio, Wav2Vec achieves an AUC of 0.600 with Random Forest, slightly outperforming prosodic features, which results in an AUC of 0.554 with Logistic Regression. CLIP visual features also perform well, producing an AUC of 0.921 with SVM. However, the best overall performance comes from hybrid fusion, combining BERT, Wav2Vec, and CLIP, where SVM reaches an AUC of 0.943, the highest among all models.

4.3 Leave-One-Video-Out Cross-Validation

To evaluate the real-world performance of our multimodal hybrid fusion model, we conducted leave-one-video-out cross-validation, iteratively excluding one video from the training set to prevent data leakage. We used only the top-performing model for both classification tasks. Results demonstrated AUC scores of 0.78 for on-task vs. off-task classification and 0.86 for sharing vs. beyond-sharing classification, with relative performance decreases of 10.86% and 8.80% compared to the baseline, respectively.

5 Discussion and Future Work

This paper explores the potential roles of multimodal machine learning models in estimating and differentiating students' various levels of SCK moment-by-moment, using noisy multimodal data collected from a real classroom while students were engaged in a project-based group activity related to AI literacy training in K-12. To our knowledge, this is one of the first studies in this particular learning context, and it also contributes to the growing body of research in multimodal machine learning and analytics to support teaching and learning in real-world educational environments.

Our study demonstrated that multimodal analysis achieved better performance in accurately estimating group-level cognitive states involved in learning, such as on-task/off-task behavior, and specific SCK-related constructs, such as sharing vs. beyond-sharing. Our findings emphasize the critical importance of verbal and non-verbal features in a complex collaborative learning environment, where students' interactions provide valuable insights into their cognitive processes conducive to learning. The result shows that body language and facial expressions are just as essential as verbal cues in assessing different levels of SCK. In addition, the results demonstrate the relationship between language and body (including the sensorimotor experiences) in SCK which are not yet extensively explored. This warrants future research on how changes in learners' body interactions with the environment result in their change of cognitive processes, such as verbal conversations.

We envision two use cases that may leverage the multimodal models, in which the model output of SCK estimates can be presented to teachers in the form of "teaching analytics" to support their pedagogical decision-making in the classroom. In **retrospective use case**, the multimodal ML model estimates the moment-by-moment construct of interest and presents the data as a multivariate time series. This view may be made available to teachers after the class session has ended, which allows them to analyze overall temporal patterns and the evolution of key metrics throughout the session to deepen their understanding of the dynamics and inform their pedagogical decision for the next session. In **Real-time use case**, the estimations are calculated in real-time, aggregated from a short time window in the recent past. This view provides teachers with "situational awareness" of the collective cognitive state of each student group, which is helpful for them to decide in real-time the type and amount of support

needed to each group. Future work will engage teachers in sense-making of the model result and co-design of the teacher-facing dashboard to ensure the design of analytics and dashboard is aligned with pedagogical practices and classroom norms, including addressing the privacy and ethical concerns from stakeholders.

6 Conclusion

With the increasing availability of multimodal sensors and advanced computational models presents new opportunities to investigate complex psychological constructs within collaborative learning environments. Drawing on a case study of AI integration in K-12 classrooms, our research provides preliminary evidence for the feasibility of extracting meaningful insights from student interactions as they engage with novel content. Future research should prioritize the integration of Multimodal Large Language Models (MLLMs) and the careful consideration of ethical implications. This work offers a foundation for advancing personalized and equitable AI literacy education in K-12 settings.

References

1. Acosta, H., Lee, S., Mott, B., Bae, H., Glazewski, K., Hmelo-Silver, C., Lester, J.: Multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning. In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 224–235 (2024)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460 (2020)
3. Boyd, R.L., Ashokkumar, A., Seraj, S., Pennebaker, J.W.: The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin **10** (2022)
4. Bradford, M., Khebour, I., Blanchard, N., Krishnaswamy, N.: Automatic detection of collaborative states in small groups using multimodal features. In: International Conference on Artificial Intelligence in Education, pp. 767–773. Springer (2023)
5. Carney, M., et al.: Teachable machine: approachable web-based tool for exploring machine learning classification. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–8 (2020). https://doi.org/10.1145/3334480.3382839
6. Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., Barro, S.: AI literacy in k-12: a systematic literature review. Int. J. STEM Educ. **10**(1) (2023).https://doi.org/10.1186/s40594-023-00418-7
7. Chango, W., Lara, J.A., Cerezo, R., Romero, C.: A review on data fusion in multimodal learning analytics and educational data mining. Wiley Interdisc. Rev. Data Min. Knowl. Disc. **12**(4), e1458 (2022)
8. Dai, Y., Lin, Z., Liu, A., Wang, W.: An embodied, analogical and disruptive approach of AI pedagogy in upper elementary education: an experimental study. Br. J. Edu. Technol. (2023). https://doi.org/10.1111/bjet.13371
9. Echeverria, V., et al.: Teamslides: a multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 112–122 (2024)

10. Eyben, F., et al.: The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016). https://doi.org/10.1109/TAFFC.2015.2457417
11. Gergen, K.J.: The Social Constructionist Movement in Modern Psychology, pp. 556–559. American Psychological Association (1992). https://doi.org/10.1037/10112-044
12. Guerin, B.: Behavior analysis and the social construction of knowledge. Am. Psychol. **47**(11), 1423–1432 (1992). https://doi.org/10.1037/0003-066X.47.11.1423
13. Gunawardena, C., Lowe, C., Anderson, T.: Analysis of a global on-line debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. J. Educ. Comput. Res. **17**(4), 395–429 (1997)
14. Gunawardena, C.N., Flor, N.V., Gomez, D., Sanchez, D.: Analyzing social construction of knowledge online by employing interaction analysis, learning analytics, and social network analysis. Q. Rev. Dist. Learn. **17**(3), 35–60 (2019)
15. Lowenfeld, V.: Creative and Mental Growth, 3rd ed. Macmillan, Oxford, England (1957)
16. Lu, W.Y., Fan, S.C.: Developing a weather prediction project-based machine learning course in facilitating AI learning among high school students. Comput. Educ. Artif. Intell. **5** (2023).https://doi.org/10.1016/j.caeai.2023.100154
17. Ng, D., Leung, J., Chu, S., Qiao, M.S.: Conceptualizing AI literacy: an exploratory review. Comput. Educ. Artif. Intell. **2**, 100041 (2021). https://doi.org/10.1016/j.caeai.2021.100041
18. Ng, D.T.K., Su, J., Leung, J.K.L., Chu, S.K.W.: Artificial intelligence (AI) literacy education in secondary schools: a review. Interact. Learn. Environ. 1–21 (2023). https://doi.org/10.1080/10494820.2023.2255228
19. Nguyen, H., Diederich, M.: Facilitating knowledge construction in informal learning: a study of tiktok scientific, educational videos. Comput. Educ. **205** (2023). https://doi.org/10.1016/j.compedu.2023.104896
20. Odiri Amatari, V.: The instructional process: a review of flanders' interaction analysis in a classroom setting. Int. J. Secondary Educ. **3**(5) (2015).https://doi.org/10.11648/j.ijsedu.20150305.11
21. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
22. Sanh, V.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
23. Ye, D., Pennisi, S.: Analysing interactions in online discussions through social network analysis. J. Comput. Assist. Learn. **38**(3), 784–796 (2022). https://doi.org/10.1111/jcal.12648
24. Zhao, L., et al.: Mets: multimodal learning analytics of embodied teamwork learning. In: LAK23: 13th International Learning Analytics and Knowledge Conference, pp. 186–196 (2023)
25. Zhou, Q., Suraworachet, W., Cukurova, M.: Harnessing transparent learning analytics for individualized support through auto-detection of engagement in face-to-face collaborative learning. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 392–403 (2024)

beSEL: A Human-Aligned LLM Teaching Assistant for Enhancing Social-Emotional Learning

Jing Zhang[1], Yazhe Niu[2(✉)], Xueyan Li[2], Peiyan Zhou[3], and Di Sun[1]

[1] Graduate School of Education, Dalian University of Technology, Dalian, China
[2] OpenDILab, Shanghai Artificial Intelligence Laboratory, ShangHai, China
{niuyazhe,lixueyan}@pjlab.org.cn
[3] School of Public Policy, University of Maryland, College Park, MD, USA

Abstract. As Social-Emotional Learning (SEL) requires personalized, context-aware interactions, educators seek more adaptive teaching support. Generative AI addresses this by providing automated and personalized assistance. However, powerful AI models like Large Language Models (LLMs) still lack the capacity for emotionally nuanced responses grounded in real-life experiences and cultural subtleties, which are essential for SEL education. Therefore, we introduce beSEL (**b**oosting **e**ducators for **s**ocial-**e**motional **l**earning), a teaching assistant system that aligns LLMs with human expertise in SEL through two core functionalities: (a) collaborates with educators to optimize lesson plans, and (b) engaging with students to support SEL knowledge and skill acquisition. We designed an efficient pipeline to construct domain-specific SEL instructional datasets via multi-round LLM-based segmentation and matching. Using these datasets, we fine-tuned Qwen2.5-3B using LoRA and evaluated the system to assess whether the generated content meets SEL teaching standards. We find that beSEL demonstrates more sufficient emotional intelligence and situational understanding than several general AI models, highlighting the effectiveness of SEL human-alignment by supervised fine-tuning.

Keywords: Social-Emotional Learning · Teaching Assistant · Generative AI · Large Language Models

1 Introduction

With an increasing emphasis on students' holistic development, Social-Emotional Learning (SEL) has become a vital part of modern education. SEL helps students develop core competencies such as self-awareness, emotional regulation, and interpersonal communication, enhancing their overall social adaptability [17]. However, traditional SEL implementation heavily relies on teacher expertise and direct interaction, yet struggles with growing student diversity challenges [3,11]. Educators face mounting pressure to simultaneously manage diverse learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 93–101, 2025.
https://doi.org/10.1007/978-3-031-99264-3_12

needs alongside varied emotional and behavioral responses. Compounded by high training demands and unequal resource distribution, this creates global shortages of qualified SEL educators, particularly in developing regions [10].

To address these challenges, generative artificial intelligence (Gen AI) has been gradually introduced into the field of education. These tools provide automated, flexible, and personalized instructional support in content generation and real-time interaction [14,18]. Previous research has shown that GenAI can help students engage in individualized and more targeted learning of SEL skills [16,19]. However, existing large language models (LLMs), such as GPT-4o, while demonstrating strong capabilities in chat and content creation, often lack sufficient emotional intelligence and situational understanding [1,6], both of which are critical in SEL education [7]. More importantly, while general LLMs adoption in education grows [12], current implementations predominantly feature student-centric or task-oriented systems, overlooking teachers' pedagogical demands for instructional flow, content curation, and classroom-context adaptation. This gap results in unclear workflow specifications and inadequate technical support for educators, limiting the effective integration of GenAI into practice.

Therefore, we present beSEL, a teaching assistant system based on supervised fine-tuned LLMs specifically for SEL instruction. The beSEL is designed to: (1) collaborate with SEL educators to enhance and refine lesson plans, and (2) engage with students to support their acquisition of SEL knowledge and skills. Due to the lack of high-quality and large-scale datasets on classroom interactions for SEL, we designed a complex and efficient data processing pipeline based on SEL classroom videos and accompanying lesson plans from elementary schools that have received awards. For the first dataset (Teachers' Response), we transcribed classroom videos, then used LLMs to iteratively segment lesson plans, aligning them with corresponding dialogues extracted from the videos to generate question-answer pairs. For the second dataset (Teaching Design), we utilized LLMs to extract instructional themes and corresponding teaching designs, forming related question-answer pairs. We trained two specialized and adaptive instructional models using LoRA-based [8] supervised fine-tuning, integrating them into the beSEL system as distinct agents: one for instructional interaction and the other for lesson plan generation. Finally, we assess beSEL's adherence to established SEL instructional standards. Our results show that beSEL consistently demonstrates emotional understanding and contextual awareness, highlighting the potential of Gen AI in SEL education.

2 Related Work

2.1 Application of GenAI in SEL

Current work on GenAI in SEL remains limited but can be broadly categorized into two areas: one examines the use of LLM to provide personalized education, such as eaSEL [16]. The other explores the application of LLMs in education based on SEL theory (PBChat [2]). While the literature identifies the potential of GenAI, there are also notable challenges and risks, particularly regarding

contextual understanding and communication [7,21]. Thus, fine-tuning LLMs is essential to ensure emotionally appropriate and educationally aligned outputs.

2.2 LLM-Based Teaching Assistants

Currently, the applications of LLM-based teaching assistants can be categorized into three main scenarios: (1) Classroom simulation refers to the use of AI agents to emulate typical classroom scenarios, such as Classroom Simulacra [23]. (2) Academic feedback refers to providing students with timely, relevant, and constructive support to address challenges during learning (CVTutor [20]). (3) Instructional resource generation refers to supporting teachers' preparation work by automating the creation of teaching materials, such as LessonPlanner [4] and SKarREC [13]. Despite their promise, most existing LLM-based teaching assistants are student-centric or task-oriented systems. Furthermore, they lack the practical, interactive assistance for effective teaching, especially in SEL. This highlights the need for a system that supports real-time classroom interaction, emotional safety, and adaptive content. In summary, unlike previous work where SEL-focused LLMs lack contextual understanding and clarity, and existing teaching assistants fail to provide practical support and genuine interaction, our work introduces a human-aligned, teacher-centered LLM system designed to address these gaps with emotionally aware and contextually relevant support.

3 Methodology and System

In this section, we propose beSEL, a human-aligned LLM teaching assistant system. Building such a system requires high-quality datasets that reflect authentic classroom interactions. However, existing SEL datasets often lack emotional nuance and contextual depth, and expert annotation remains costly and labor-intensive. To address these challenges, we designed a pipeline to construct datasets, as described in Sect. 3.1. The fine-tuning process is detailed in Sect. 3.2, and the overview of beSEL system is introduced in Sect. 3.3.

3.1 Dataset

We collected 362 SEL classroom videos and corresponding lesson plans from elementary schools that have received awards and established an effective processing pipeline to generate two specialized datasets, as shown in Fig. 1. All classroom videos and lesson plans were collected from publicly accessible sources or from schools that granted explicit permission for research use. No identifiable information about students or teachers was included in the final dataset. All data were anonymized and used in accordance with ethical research standards.

The first dataset is *Teachers' Response*. This dataset was constructed to capture authentic instructional interactions and pedagogical strategies. First, we used FunASR [5] to transcribe videos and distinguish speakers. Second,

Qwen2.5-72B-Instruct [24] was used to break down the lesson plans into activities, based on their instructional design. Third, we applied Qwen2.5-72B-Instruct to match the processed video transcripts with the segmented teaching activities. We found that directly prompting an LLM to perform segmentation often resulted in coarse-grained outputs with vague semantic boundaries, making it difficult to align with specific instructional behaviors. Therefore, we designed a multi-stage segmentation process using a chain-of-prompt strategy, where the LLM first segments content by broad instructional phases and then refines it into detailed behavioral units like guided questioning and emotional feedback. Lastly, the LLM generated questions from the instructional context and retrieved corresponding teacher utterances from the video transcripts as answers.

The second dataset is *Teaching Design*. This dataset was constructed to capture expert-level teaching design. First, LLM is applied to extract topics from the lesson plans and generate questions. The LLMs then extract the teaching design, which is paired with the questions to form a Q&A dataset.

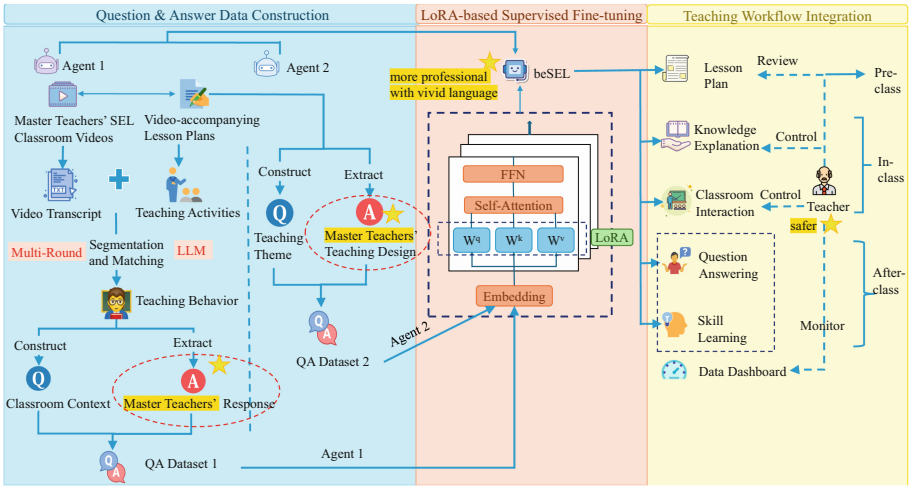


Fig. 1. Overview of the beSEL system. It includes: (1) a multi-round data construction pipeline to generate master teacher-based Q&A datasets; (2) LoRA-based fine-tuning of two instructional agents; and (3) integration into a teacher-centered SEL workflow with vivid outputs, enhanced safety, and full-process instructional support.

3.2 Supervised Fine-Tuning

A LLM can be fine-tuned by adjusting it to make it more effective for specific purpose. We used Qwen2.5-3B [24] as the baseline model and applied LoRA fine-tuning on two distinct datasets to obtain task-specific models. To preserve the generalization of the model and the retention of knowledge, we also introduced a KL divergence constraint with the base model during fine-tuning. LoRA

fine-tuning works by introducing low-rank matrices into the attention layers of the model, allowing the model to adapt to tasks with fewer trainable parameters while maintaining performance, thus reducing computational burdens [8]. This approach has been shown to be resource-efficient while optimizing large models [15]. To validate the effectiveness of fine-tuning, we invited SEL teachers to assess whether generated content meets SEL teaching standards.

3.3 BeSEL System

Based on constructed datasets and fine-tuned models, we developed an intelligent teaching assistant system–beSEL. This system is a teacher-centered, human-aligned LLM assistant that supports a range of automated instructional tasks and integrates efficiently into teachers' workflows, as shown in Fig. 1.

In the pre-class phase, beSEL assists teachers in optimizing lesson plans and enables them to review and edit LLM-generated content to ensure instructional quality and safety. During the in-class phase, beSEL supports knowledge explanation and classroom interaction under teacher supervision, delivering real-time content while fostering dynamic exchanges and emotional connection. In the after-class phase, beSEL supports students through question answering to reinforce understanding, and incorporates role-playing and scenario-based dialogues to enhance SEL skill learning. In addition, beSEL integrates a data dashboard module that visualizes the data of after-class interaction between students. It

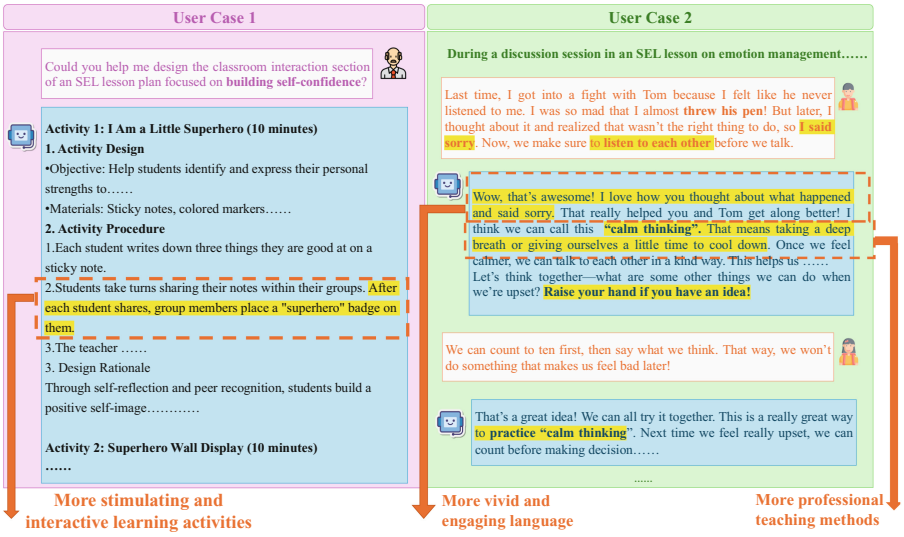


Fig. 2. User cases of the beSEL system. beSEL facilitates interactive lesson design (left) and SEL-related discussions (right), demonstrating its capability to generate more stimulating activities, more vivid and engaging language, and more professional teaching methods tailored to SEL instruction.

monitors students' emotions and learning in real time and flags high-risk language for timely teacher intervention.

4 Preliminary Results and Discussion

We conducted a pilot study with 39 SEL teachers from an elementary school to evaluate the effectiveness of our fine-tuned models, Fig. 2 shows user cases. Teachers were then asked to complete evaluation questionnaires based on the Washington SEL Implementation Brief: For Educators [22], which outlines core SEL teaching competencies. All participating teachers and students provided informed consent prior to the study. These were distilled into five dimensions: Inclusive Learning Climate, Emotional Regulation Support, Cultural & Equity Awareness, Social Connection Promotion, and Goal & Efficacy Support. Two tailored questionnaires were developed to assess each agent across these dimensions, using a 5-point Likert scale from -2 to +2, where 0 indicates a practical baseline. Since GPT-4o mini is similar in size to our fine-tuned models, teachers also evaluated Qwen2.5-3B and GPT-4o mini following the same procedure. Results showed that beSEL achieved the highest mean scores in both instructional interaction (mean = 1.092) and lesson plan generation (mean = 1.008).

Table 1. Paired Samples t-test Results for Mean Difference. Task 1 is instructional interaction, Task 2 is lesson plan generation. Exp. 1 represents beSEL, Exp. 2 represents Qwen2.5-3B, and Exp. 3 represents GPT-4o mini. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Dimension	Exp.1 VS. Exp.2		Exp.1 VS. Exp.3	
	Task 1	Task 2	Task 1	Task 2
Inclusive Learning Climate	0.564***	0.397***	0.295***	0.244**
Emotional Regulation Support	0.641***	0.474***	0.154*	0.321***
Cultural & Equity Awareness	0.641***	0.449***	0.282***	0.205**
Social Connection Promotion	0.538***	0.410***	0.231**	0.026
Goal & Efficacy Support	0.590***	0.474***	0.141	0.128
Total	0.595***	0.441***	0.221**	0.185**

To test significance, paired samples t-tests were conducted across all dimensions and tasks, as shown in Table 1. The results demonstrate that in Emotional Regulation Support, beSEL exceeded Qwen2.5-3B by a mean difference of 0.641 and GPT-4o mini by 0.474 in instructional interaction (p ¡ 0.01), indicating its superior capacity to recognize and respond to students' emotional states. In Cultural & Equity Awareness, the differences reached 0.641 and 0.449 (p ¡ 0.01), suggesting that beSEL is more effective in promoting inclusive, culturally responsive teaching practices. Notably, beSEL achieved significant gains in all five dimensions compared to Qwen2.5-3B (p ¡ 0.01), and showed significant differences in most dimensions when compared to GPT-4o mini. This suggests

that fine-tuning beSEL led to substantial improvements, enabling the model to exhibit more sufficient emotional intelligence and situational understanding for SEL-related teaching tasks, compared to general-purpose models in both lesson preparation and real-time classroom interaction.

To address cultural bias and promote fairness in SEL applications, we examine beSEL's transferability across diverse cultural and educational settings. Given regional and age-specific differences in SEL practices, we propose three adaptation strategies: (1) embedding culturally grounded examples in prompts to reflect local norms and values; (2) employing modular plugins–such as web search–to tailor SEL standards and feedback to institutional and regional contexts; and (3) integrating lightweight multi-task fine-tuning with hierarchical prompt templates to better capture age-specific emotional and linguistic patterns. These strategies support beSEL's real-world adaptability.

5 Conclusion and Future Work

This study introduces beSEL, a human-aligned LLM teaching assistant that supports the entire SEL instructional process. Through customized instructional agents and a teacher-centered design, beSEL demonstrates the potential of Gen AI in enhancing emotional awareness, situational understanding, and professional classroom interaction. In future work, we plan to extend beSEL to multimodal models like Step-Audio [9], to support tone recognition and speech-based instruction for more expressive classroom interactions. We also aim to deploy the system in real-world classroom settings, collecting both positive and negative feedback from teachers and students. By incorporating Reinforcement Learning from Human Feedback (RLHF), we will design a pipeline that quickly adapts the model to specific teaching preferences and classroom contexts. Furthermore, we will explore fairness-aware alignment strategies to ensure equitable and responsible deployment of beSEL across diverse educational environments.

Acknowledgments. This study was funded by China's Ministry of Education of Humanities and Social Science project (Grant No. 23YJA880044); Fundamental Research Funds for the Central Universities (Dalian University of Technology) (Grant No. DUT24LAB130).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amin, M.M., Mao, R., Cambria, E., Schuller, B.W.: A wide evaluation of ChatGPT on affective computing tasks. IEEE Trans. Affect. Comput. (2024)
2. Chen, P., Fan, Z., Lu, Y., Xu, Q.: Pbchat: enhance student's problem behavior diagnosis with large language model. In: International Conference on Artificial Intelligence in Education, pp. 32–45. Springer (2024)

3. Durlak, J.A., Weissberg, R.P., Dymnicki, A.B., Taylor, R.D., Schellinger, K.B.: The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. Child Dev. **82**(1), 405–432 (2011)
4. Fan, H., Chen, G., Wang, X., Peng, Z.: Lessonplanner: assisting novice teachers to prepare pedagogy-driven lesson plans with large language models. In: Proceedings of the 37th Annual ACM Symposium on UIST, pp. 1–20 (2024)
5. Gao, Z., et al.: Funasr: a fundamental end-to-end speech recognition toolkit. arXiv preprint arXiv:2305.11013 (2023)
6. Havaldar, S., Rai, S., Singhal, B., Liu, L., Guntuku, S.C., Ungar, L.: Multilingual language models are not multicultural: a case study in emotion. arXiv preprint arXiv:2307.01370 (2023)
7. Henriksen, D., Creely, E., Gruber, N., Leahy, S.: Social-emotional learning and generative AI: a critical literature review and framework for teacher education. J. Teacher Educ. 00224871251325058 (2025)
8. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
9. Huang, A., et al.: Step-audio: unified understanding and generation in intelligent speech interaction. arXiv preprint arXiv:2502.11946 (2025)
10. Jagers, R.J., Rivas-Drake, D., Williams, B.: Transformative social and emotional learning (SEL): Toward SEL in service of educational equity and excellence. Educ. Psychol. **54**(3), 162–184 (2019)
11. Jones, S.M., Kahn, J.: The evidence base for how we learn: supporting students' social, emotional, and academic development. Consensus statements of evidence from the council of distinguished scientists. Aspen Institute (2017)
12. Kasneci, E., et al.: ChatGPT for good? on opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
13. Li, Q., et al.: Learning structure and knowledge aware representation with large language models for concept recommendation. arXiv preprint arXiv:2405.12442 (2024)
14. Li, X., Chen, X., Niu, Y., Hu, S., Liu, Y.: PsyDI: towards a personalized and progressively in-depth chatbot for psychological measurements. arXiv preprint arXiv:2408.03337 (2024)
15. Liao, X., Wang, C., Zhou, S., Hu, J., Zheng, H., Gao, J.: Dynamic adaptation of LoRA fine-tuning for efficient and task-specific optimization of large language models. arXiv preprint arXiv:2501.14859 (2025)
16. Shen, J., Chen, J.K., Findlater, L., Smith, G.D.: eaSEL: promoting social-emotional learning and parent-child interaction through AI-mediated content consumption. arXiv preprint arXiv:2501.17819 (2025)
17. Shi, J., Cheung, A.C.: Effective components of social emotional learning programs: a meta-analysis. J. Youth Adolesc. **53**(4), 755–771 (2024)
18. Tan, K., Pang, T., Fan, C., Yu, S.: Towards applying powerful large AI models in classroom teaching: opportunities, challenges and prospects. arXiv preprint arXiv:2305.03433 (2023)
19. Tang, Y., et al.: Emoeden: applying generative artificial intelligence to emotional learning for children with high-function autism. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–20 (2024)
20. Teng, D., et al.: Investigating the utilization and impact of large language model-based intelligent teaching assistants in flipped classrooms. Educ. Inf. Technol. 1–34 (2024)
21. Warr, M., Oster, N.J., Isaac, R.: Implicit bias in large language models: experimental proof and implications for education. J. Res. Technol. Educ. 1–24 (2024)

22. Washington Office of Superintendent of Public Instruction: Washington SEL implementation brief: For educators (2023). https://ospi.k12.wa.us/sites/default/files/2024-02/washington_sel_implementation_brief_educators122023.pdf. Accessed 03 Apr 2024
23. Xu, S., Wen, H.N., Pan, H., Dominguez, D., Hu, D., Zhang, X.: Classroom simulacra: building contextual student generative agents in online education for learning behavioral simulation. arXiv preprint arXiv:2502.02780 (2025)
24. Yang, A., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)



Automated Essay Scoring Using Logical Structure Graph Information

Yoshihiro Kato(✉)

Benesse Educational Research and Development Institute, Tokyo, Japan
y-kato@mail.benesse.co.jp

Abstract. This paper proposes an automated essay scoring (AES) model that utilizes the logical structure of an essay. We utilize the entire sentence containing the logical elements to estimate the relationships between the sentences. Then, the proposed method extracts relevant features using graph attention networks that can consider the graph structure and combines with the conventional AES model. Experiments conducted on the ASAP dataset demonstrated that the proposed model obtains higher accuracy than the baseline and conventional methods. Furthermore, it achieved higher accuracy even in cases of random or absent argumentative structures, as well as those estimated by LLMs.

Keywords: Automated Essay Scoring · Argument Mining · Graph Neural Networks

1 Introduction

Essay writing, which requires learners to express their own thoughts, has long been used in learning materials and tests. However, essay scoring requires not only the labor costs of the graders but also the accuracy, consistency, and reliability of the grading. Thus, computer-based AES has been performed for many years to improve the consistency and reliability of the scoring results, reduce the burden of scoring tasks, lower costs, reduce essay scoring time and facilitate the early return of the scoring results [8].

In this study, we attempt to improve the accuracy by incorporating the features of important scoring criteria into an AES model based on a language model. We focus on AES models that utilize information about the logical structure. The logical structure influences a trait score of organization and holistic score, and it can be estimated using the argument mining method. Nguyen et al. [6] proposed an AES model that utilizes a statistical count of logical elements of essays, e.g., number of claims and premises. In addition, Yamaura et al. [10] improved accuracy by combining the output of a language model with the output of a model that incorporates logical structure information into an attention mechanism of the language model. However, these models convert the argumentative structure or elements into frequency and attention to fit the AES models.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 102–109, 2025.
https://doi.org/10.1007/978-3-031-99264-3_13

Thus, we propose an AES model that employs graph attention networks (GAT) to extract the graph information of the logical structure and concatenate it with an AES model based language model.

The accuracy of the proposed model was validated on benchmark data, and the results demonstrated that the proposed model achieved higher accuracy than the baseline method, conventional methods and other similar models. We found that the proposed model obtained higher accuracy, thereby confirming its effectiveness.

2 Graph Attention Networks

Graph Neural Networks (GNNs) are a neural network method to handle graph–structured data [2]. GNNs aggregate the features of nodes in a graph along with the information from neighboring nodes connected by edges, transforming the graph into a vector representation. Given a graph $\mathcal{G} \in (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V} = \{1, \ldots, i, \ldots, n\}$ and edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, where an edge from node j to node i is denoted $(j, i) \in \mathcal{E}$, the set of node features $\{\mathbf{h}_i \in \mathbb{R}^d \mid i \in \mathcal{V}\}$ and the edge set $\mathcal{E}$ serve as inputs to the GNN. The GNN updates the node features without altering the graph structure, and the updated node features $\{\mathbf{h}'_i \in \mathbb{R}^{d'} \mid i \in \mathcal{V}\}$ are computed by aggregating the neighboring nodes $\mathcal{N}_i = \{j \in \mathcal{V} \mid (j, i) \in \mathcal{E}\}$ using an aggregation function AGGREGATE$(\cdot)$ and a function f. Here, AGGREGATE combines the features of the neighboring nodes, and f updates the node features based on the aggregated information. This process allows the GNN to capture the structural information of the graph and incorporate it into the node feature updates, which can then be utilized in an AES.

$$\mathbf{h}'_i = f_\theta(\mathbf{h}_i, \text{AGGREGATE}(\{\mathbf{h}_j \mid j \in \mathbf{N}_i\})) \tag{1}$$

Many GNNs have been proposed based on the differences in the f and AGGREGATE. The function f can be a linear transformation or an activation function, e.g., ELU, while AGGREGATE can involve different operations, e.g., mean or max pooling.

Recently, Graph Attention Networks (GATs) [1], which apply the attention mechanisms used in language models to aggregate neighboring nodes, have become increasingly popular. The attention method calculates weights for each element of the input and extracts important information by taking the weighted average of these values, where the evaluation function $e : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ represents the weight of the edge (j, i) from nodes j to i.

Here, $\mathbf{a} \in \mathbb{R}^{2d'}$ and $\mathbf{W} \in \mathbb{R}^{d' \times d}$ are learned from the data, and $\|$ denotes the concatenation of the vectors. By normalizing over all neighboring nodes $j \in \mathcal{N}_i$ using the SoftMax function, the attention α_{ij} can be expressed as follows.

$$\alpha_{ij} = \frac{\exp(\mathbf{a}^\top \text{LeakyReLU}\left(\mathbf{W} \cdot [\mathbf{h}_i \| \mathbf{h}_j]\right))}{\sum_{j' \in \mathcal{N}_i} \exp\left((\mathbf{a}^\top \text{LeakyReLU}\left(\mathbf{W} \cdot [\mathbf{h}_i \| \mathbf{h}_{j'}]\right)\right)} \tag{2}$$

Then, GAT computes a weighted average of the transformed features of the neighbor nodes, using the normalized attention coefficients:

$$\mathbf{h}_i' = f_\theta(\mathrm{GAT}(\mathcal{G})) = f_\theta\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij} \cdot \mathbf{W}\mathbf{h}_j\right) \tag{3}$$

3 Proposed Method

3.1 BERT+GAT

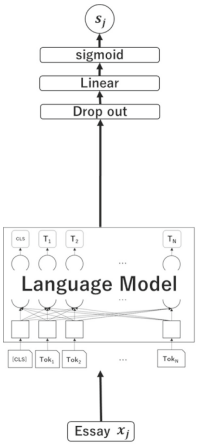


Fig. 1. Architecture of Basic model

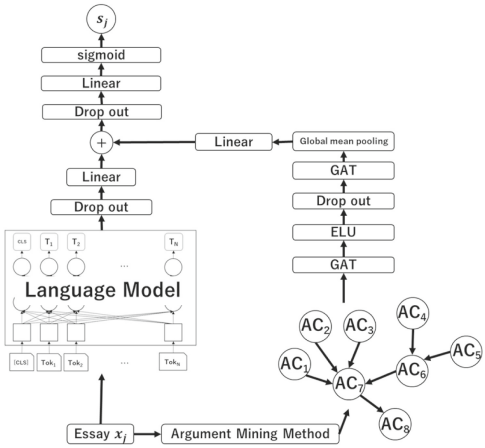


Fig. 2. Architecture of Proposed model

In this study, we employed the Bi-Directional Encoder Representations from Transformers (BERT) as the AES model, due to its use in many previous studies. The BERT is constructed by stacking multiple Transformers (12 layers for the base model and 24 layers for the large model; we use the base model). Figure 1 shows the architecture of the AES Model using BERT. Here, the j-th essay x_j is represented as $x_j = \{w_{j0}, \ldots, w_{jn_j}\}$, where w_{ji} denotes the i-th token of the j-th essay, and n_j denotes the total number of tokens in the j-th essay. Let s_j be the score of the j-th essay, $\mathbf{h}_{[\mathbf{CLS},\mathbf{j}]}$ be the context vector from the BERT model (the [CLS] classification token is added to identify the beginning of the sentence), N be the total number of essays, Linear$(\cdot)$ is a fully-connected layer, and $\sigma(\cdot)$ be the sigmoid function. In addition, the mean squared error (MSE) between the true score s_j and the predicted score $\hat{s}_j$ is used as the evaluation function (i.e., loss function), and the AES model is optimized by minimizing the loss function.

$$\mathrm{argmin}\, \frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s}_i)^2 \tag{4}$$

$$s_j = \sigma\left(\mathrm{Linear}\left(\mathbf{h}_{[\mathbf{CLS}],\mathbf{j}}\right)\right) \tag{5}$$

Figure 2 shows the architecture of the proposed model, which can handle the argumentative structure graphs of essays using two GATs. A argumentative structure graph of j-th essay $\mathcal{G}_j$, estimated through argument mining, is input into the first GAT ($\mathrm{GAT}^{(0)}$), which applies the ELU activation function, and its output is then input into the second GAT($\mathrm{GAT}^{(1)}$), from which $\mathbf{h}_{[\mathbf{GAT}],\mathbf{j}}$ is obtained through global max pooling. The proposed BERT+GAT model concatenates the feature vectors from BERT and GATs after linear transformations, which is followed by an additional linear transformation.

$$s_j = \sigma\left(\mathrm{Linear}\left(\mathrm{Linear}\left(\mathbf{h}_{[\mathbf{CLS}],\mathbf{j}}\right) + \mathrm{Linear}\left(\mathbf{h}_{[\mathbf{GAT}],\mathbf{j}}\right)\right)\right) \tag{6}$$

$$\mathbf{h}_{[\mathbf{GAT}],\mathbf{j}} = \frac{\mathbf{1}}{\mid \mathcal{V} \mid} \sum_{\mathcal{V}} \mathrm{GAT}^{(\mathbf{1})}(\mathrm{ELU}(\mathrm{GAT}^{(\mathbf{0})}(\mathcal{G}_{\mathbf{j}}))) \tag{7}$$

3.2 Argument Mining for AES

Argument mining involves the automatic extraction of the argumentative structure from the text data. Generally, the argumentative structure is estimated by using the following process. (1) Argument Component Identification (ACI): Extracting the elements of argumentative structure, such as clauses or phrases, which are referred to as Argument Components (ACs). (2) Argument Component Classification (ACC): Classifying the elements into categories such as Major Claim, Claim, and Premise. (3) Argument Relation Classification (ARC): Drawing directed edges between elements to indicate that one element support or attack another element, while "neutral" does not have any edges. (4) Argument Structure Estimation (ASE): Estimating argumentative structure based on the results of the ACC and ARC. In addition, methods that treat these steps as a single task and directly output an argumentative structure from textual data have also been proposed [5]. Morio et al. proposed a MT-AM approach that performs ACI, ACC and ARC using Longformer, which can handle longer texts. We adopted an argumentative structure estimation method based on the MT-AM approach.

When applying the MT-AM approach to the experimental data, it was observed that clauses and phrases, symbols and meaningless sequences of one or two words were all extracted as ACs, which likely occurred because the data used in the experiments lack ground truth labels for argumentative structures, which prevents fine-tuning of the MT–AM approach on the target data. In addition, essays written by students frequently lack coherent argumentative structures and tend to contain numerous vocabulary and grammatical errors. When these results were directly input into the proposed method, it showed an improvement over the baseline; while, it performed worse than conventional methods [10]. Thus, we propose an argumentative structure method for AES based on the inference results of the MT-AM approach. Specifically, we propose sentArg, in which a sentence containing ACs inferred by the MT-AM approach is treated as an AC. Note that the classification of relations between argument elements by the MT-AM approach is not a sentence; thus, we need to estimate a relation

for between sentences. In addition, we used LogBERT [3] for the relationship classification of the sentence. Yohan et al. classified the relationships between sentences into four logical relations (factual consistency, emotional consistency, causal relations, and normative relations), and they formulated them using probabilistic soft logic. The LogBERT uses these results to train a classification model using BERT, and optimized it on a large-scale discussion dataset. The LogBERT classifies the relationship between the sentences as either support, attack, or neutral.

After performing relation classification, we determined the structure using integer linear programming (ILP). Here, we explain the ILP proposed by Stab et al. [7]. Stab et al. set constraints to determine the structure based on logical consistency and ease of interpretation, as represented by Equations (8) to (14).

$$\underset{x}{\text{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij} \tag{8}$$

s.t.

$$\forall_i : \sum_{j=1}^{n} x_{ij} \leq 1 \tag{9}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \leq n - 1 \tag{10}$$

$$\forall_i : x_{ii} = 0 \tag{11}$$

$$\forall_i \forall_j : x_{ij} - a_{ij} \leq 1 \tag{12}$$

$$\forall_i \forall_j : a_{ik} - a_{ij} - a_{jk} \leq -1 \tag{13}$$

$$\forall_i : a_{ii} = 0 \tag{14}$$

The presence of an edge between nodes is indicated by x_{ij}, where $x_{ij} = 1$ if there is a logical relationship from node i to node j, and $x_{ij} = 0$ if there is no relationship. The weight w_{ij} represents the predicted probability from the relation classification step. Equations (9) and (10) impose the constraint that each node can have at most one outgoing edge. Equation (11) ensures that no node has an edge to itself. The a_{ij} is 1 if node i can reach node j, and 0 otherwise. Equations (12) to (14) are constraints to prevent cyclic structures. The structure was estimated by imposing constraints that prevent the edges from being drawn between the same nodes and from one node to multiple nodes.

4 Experiment and Result

In the experiments conducted in this study, the publicly available dataset from the Automated Student Assessment Prize (ASAP) was used to evaluate the proposed method. The ASAP dataset has been widely used for AES research [4,6,9,10]. We employed the quadratic weighted kappa (QWK) to evaluate the performance of the model. The ASAP dataset and the QWK have been widely

used for AES research. The experiment was conducted using 5-fold cross validation. In addition, the model was trained using the AdamW optimizer to minimize MSE loss function over the development data, and training was stopped when the loss of the development data did not improve five consecutive epochs. We used dropout regularization to prevent overfitting, setting dropout probabilities from BERT to linear and from GAT to linear at 0.5.

We compared the proposed models with the BERT, EASE+Arg, BERT+LS, NPCR, chatGPT MTS models. The NPCR [9] achieves high accuracy by predicting the score difference between the target essay and the learning data, and then averaging the prediction results by changing the pairs multiple times for the target essays. Lee et al. [4] proposed the multi trait specialization (MTS), which extracts key points from the scoring criteria used by human evaluators to generate new scoring criteria, which ensures consistent scoring by ChatGPT using MTS (i.e., chatGPT MTS).

To demonstrate the effectiveness of sentArg, we also considered cases where the ACs were conventional word strings, e.g., clauses and phrases. Note that only–mt–am uses only the information of the inference results of the MT–AM approach without applying RC and ILP. The method of applying relational classification and ILP as in sentArg to logical elements of mt–am is called phraseArg. We also verified a method that does not use ILP after relationship classification (referred to as notILP).

To the best of our knowledge, there are no publicly available datasets with correct labels for argumentative structures that can be used to validate AES. Thus, it is difficult to calculate the estimation accuracy of the argumentative structures and verify the effectiveness of the BERT+GAT model that inputs argumentative structure estimation method and its results. In contrast to the methods reported by previous studies [5,7], sentArg treats the ACs as sentences, which makes it difficult to validate accuracy using data from previous studies. Therefore, in the current study, we compared the results of inputting randomly determined graphs (random). If BERT+GAT with sentArg achieves higher accuracy than inputting the randomly determined graphs, the proposed method is said to be effective.

Considering the remarkable progress of LLMs, we also examined their use for inferring argumentative structures using chatGPT. In this study, we employed zero-shot learning with chatGPT–4o mini for the ASAP dataset without correct labels.

Table 1 shows the average QWK values of 5-fold cross validation for each problem ID, where Avg.(1–8) represents the overall average score. The bold text in the table indicates the top accuracy among all models, the underlined text indicates the top accuracy among the BERT+GAT and the wavy underline indicates the top accuracy within DeBERTa. In the BERT+GAT (sentArg) models, the average accuracy of sentArg is higher than that of the other BERT+GAT models. Based on the results, which demonstrate that sentArg and phraseArg achieved higher accuracy than the random methods, it is effective to input the logical structure into the GAT. In addition, the BERT+GAT (sentArg) and

Table 1. QWK evaluation scores obtained on ASAP dataset

Model	Problem ID								Avg.
	1	2	3	4	5	6	7	8	(1–8)
BERT	0.814	0.675	0.649	0.814	0.795	0.811	0.825	0.713	0.762
EASE+Arg [6]	0.832	0.689	–	–	–	–	–	–	–
BERT+LS [10]	0.815	0.672	0.693	0.816	0.809	0.814	0.829	0.717	0.771
NPCR [9]	**0.856**	**0.750**	**0.756**	0.851	**0.847**	**0.858**	0.838	**0.779**	**0.817**
chatGPT–MTS [4]	0.138	0.443	0.502	0.611	0.662	0.668	0.261	0.157	0.430
BERT+GAT(sentArg)	0.818	0.703	0.689	0.818	0.812	0.820	0.834	0.705	0.775
BERT+GAT(sentArg,notILP)	0.810	0.670	0.679	0.805	0.796	0.814	0.830	0.723	0.766
BERT+GAT(phraseArg)	0.809	0.698	0.685	0.819	0.805	0.811	0.836	0.709	0.771
BERT+GAT(phraseArg,notILP)	0.802	0.676	0.688	0.805	0.808	0.814	0.828	0.678	0.762
BERT+GAT(only–mt–am)	0.810	0.680	0.702	0.818	0.805	0.801	0.834	0.685	0.767
BERT+GAT(random)	0.803	0.676	0.692	0.815	0.802	0.808	0.829	0.696	0.765
BERT+GAT(chatGPT)	0.828	0.670	0.689	0.822	0.799	0.803	0.833	0.707	0.769
DeBERTa	0.813	0.702	0.680	0.798	0.809	0.824	0.824	0.749	0.775
DeBERTa+GAT(sentArg)	0.808	0.666	0.700	0.830	0.826	0.831	0.838	0.719	0.777
DeBERTa+GAT(phraseArg)	0.819	0.683	0.706	0.800	0.819	0.831	0.836	0.699	0.774

phraseArg methods outperformed the notILP and only-mt-am; thus, it is better to use the constrained structure proposed in this method rather than utilizing the results from the AM method directly. Furthermore, the higher accuracy of sentArg compared with phraseArg indicates that the proposed method, which treats sentences as ACs rather than adopting the savings approach of AM methods, is effective. We also compared the proposed models with the EASE+Arg and BERT+LS, which also consider the features of the argumentative structures, which is similar to our models. For problem 2, the BERT+GAT (sentArg) performed better than EASE+Arg. In addition, for all problems (except problems s 3 and 8), the BERT+GAT (sentArg) outperformed the BERT+LS and achieved a higher of Avg.(1–8). Considering these factors, it is evident that methods like the proposed approaches that treat the sentence as AC and comprehensively estimate the RC are preferable when considering the argumentative structure. DeBERTa+GAT (sentArg) achieved better results than the DeBERTa and DeBERTa+GAT (phraseArg), which indicates that the proposed method remains effective even with an increase in the parameters of language models.

5 Conclusion

This paper has proposed an AES model that considers the argumentative structure graph of essays and demonstrated the effectiveness of the proposed method experimentally on a benchmark dataset. The proposed method treats sentences

as ACs, unlike conventional studies of AM that treat sequences of tokens, e.g., phrases and clauses, as ACs. In cases where the argumentative structure is random or where no logical structure exists in the essay, or where the argumentative structures estimated by LLMs, we validated the argumentative structures and demonstrated that the proposed method, specifically BERT+GAT(sentArg), achieved higher accuracy, thereby demonstrating its effectiveness. Although the proposed method did not reach the performance of the state-of-the-art models, it achieved more accurate results than the similar BERT+LS and EASE+Arg.

References

1. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? arXiv preprint arXiv:2105.14491 (2021)
2. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings. In: 2005 IEEE International Joint Conference on Neural Networks, vol. 2, pp. 729–734. IEEE (2005)
3. Jo, Y., Bang, S., Reed, C., Hovy, E.: Classifying argumentative relations using logical mechanisms and argumentation schemes. Trans. Assoc. Comput. Linguist. **9**, 721–739 (2021)
4. Lee, S., Cai, Y., Meng, D., Wang, Z., Wu, Y.: Unleashing large language models' proficiency in zero-shot essay scoring. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 181–198 (2024)
5. Morio, G., Ozaki, H., Morishita, T., Yanai, K.: End-to-end argument mining with cross-corpora multi-task learning. Trans. Assoc. Comput. Linguist. **10**, 639–658 (2022)
6. Nguyen, H., Litman, D.: Argument mining for improving the automated scoring of persuasive essays. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
7. Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Comput. Linguist. **43**(3), 619–659 (2017)
8. Uto, M.: A review of deep-neural automated essay scoring models. Behaviormetrika **48**(2), 459–484 (2021). https://doi.org/10.1007/s41237-021-00142-y
9. Xie, J., Cai, K., Kong, L., Zhou, J., Qu, W.: Automated essay scoring via pairwise contrastive regression. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 2724–2733 (2022)
10. Yamaura, M., Fukuda, I., Uto, M.: Neural automated essay scoring considering logical structure. In: International Conference on Artificial Intelligence in Education, pp. 267–278. Springer (2023)



Leveraging AI for Persons with Disabilities: Investigating Their Needs and Gaps in AI-Supported Assistance

Rolf Feichtenbeiner(✉), Lorenz Reichert, and Susan Beudt(✉)

German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
{rolf.feichtenbeiner,lorenz.reichert,susan.beudt}@dfki.de

Abstract. Persons with disabilities are still facing widespread inequalities in the areas of education and work. AI-supported technologies show great potential in supporting persons with disabilities in vocational education and training, vocational rehabilitation, and the workplace. This paper examines the current state of developments in AI-supported assistance systems, using the example of persons with hearing impairments and visual impairments. Findings on their support needs are presented and compared to existing functions of AI assistance that were either designed to meet those specific needs or display great potential to support them in educational and workplace settings. The goal is to examine research and development gaps regarding AI assistance for persons with different kinds of disabilities. We provide first insights indicating which of their needs are currently underrepresented or unmet.

Keywords: Accessibility · AI-supported Assistance · Assistive Technology · Research and Development Gaps · Learning on the Job · Vocational Training

1 Introduction

As AI increasingly impacts all parts of our lives posing great potential but also risks for education and the world of work, this development plays an even more essential role for underrepresented groups, such as persons with disabilities. According to the last UN SDG progress report [22], considerable action is still required concerning equal access to education and work for marginalized groups worldwide. AI may present further challenges regarding equity and equality in both areas. A lack of digital literacy (as such a prerequisite for AI literacy; [10]) seems to predict the digital divide, and persons with disabilities seem to be affected differently by a disability digital divide [20]. Great efforts should be made to counter the risk that AI will increase the disability digital divide due to its complexity and lack of explainability for a diverse user group [3]. Persons with disabilities should be more substantially considered (e.g., representation in training data, [16]; prevention of disability bias, e.g. [26]) and included in the future design and development of AI technologies. Further, existing AI technologies, e.g., assistance technologies, can significantly contribute to greater participation in educational processes and the labor

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 110–117, 2025.
https://doi.org/10.1007/978-3-031-99264-3_14

market [16]. For a long time, there has been no systematic overview of AI technologies that can support persons with disabilities in education and work. In a previous project[1], we researched AI-supported assistance technologies (in the following, abbreviated as 'AI assistance') and analyzed them concerning various criteria to identify appropriate AI assistance and suitable use cases in vocational rehabilitation and workplace settings. AI assistance used in vocational rehabilitation is tied closely to learning processes as vocational rehabilitation is an important educational system that supports persons with disabilities to qualify for integration into the labor market.

We had already found initial indications that there is more AI assistance available for certain types of disabilities than for others and that specific gaps in offerings may exist. In the current project[2], we wanted to investigate this further and determine gaps in research and development of AI assistance for support in work and learning for persons with various types of disabilities (i.e., visual impairments, hearing impairments, intellectual and learning disabilities, and mental disabilities). One of the objectives is to establish a match between the needs of users with disabilities and existing AI assistance. This paper presents first selected results of the work on the support needs of persons with visual impairments and persons with hearing impairments. Analyzing research and development gaps in this field should provide valuable impetus for future AI research and development.

2 Methodology

As the study design is person-centered, the first step is understanding the barriers persons with disabilities face in their work lives and educational processes. Disabilities vary both between groups of persons with different forms of disabilities and within these groups. By focusing on different types of disabilities, more precise statements about the need for AI research and development can be derived. AI-supported assistance technologies and their functions are analyzed in terms of their potential to overcome these barriers. Finally, barriers and technologies were compared to identify research and development gaps. This study focused on barriers and AI assistance in work life, but many barriers and AI technologies are also relevant for educational processes.

Data Collection and Analysis

Four focal points were selected for analysis (guided by the number of individuals and the status and availability of AI assistance for occupational participation according to a previous project): persons with visual impairment, hearing impairment, intellectual and learning disabilities, and mental disabilities. In this article, we present an analysis of interim results, focusing on two of the four target groups: persons with visual impairment or blindness and persons with hearing impairment or deafness.

Barrier and Needs Assessment. Relevant barriers were determined through a secondary analysis of Design Thinking workshop results derived from the predecessor

[1] KI.ASSIST, funded by the Federal Ministry of Labour and Social Affairs (BMAS).

[2] KI-Kompass Inklusiv, funded by the Federal Ministry of Labour and Social Affairs (BMAS).

project KI.ASSIST (2019–2022). The digital workshops were delivered in one professional rehabilitation institution (n = 10) and one company (n = 10) and identified personas and their central problems and needs in the vocational rehabilitation and world of work (for a description of the method see [7, 21]). In addition, a narrative literature review was conducted (June–August 2023) to gain a better understanding of barriers encountered. Publications found via Web of Science, Scopus, and PsycINFO using predefined search terms were included if they a) addressed either work and learning or vocational education and training settings, b) identified barriers for persons with disabilities as research findings and c) were empirical studies. The review includes overview articles published from 2017 to 2023 and individual publications from 2000 to 2023. Next, disability-specific focus groups will take place in 2025 with persons with disabilities and AI researchers to validate the results and identify research and development needs.

Needs Analysis. All barriers from the workshops and the narrative literature review were collected. The International Classification of Functioning, Disability and Health (ICF) (WHO, 2001) chapters of activities and participation served as the initial framework for categorization. Two independent raters assigned barriers to these chapters if they matched according to the descriptions or were equivalent to a subordinate second-level ICF item. For instance, a relevant chapter was learning and applying knowledge, with corresponding second-level items d166 reading and d170 writing. However, as relevant discrepancies emerged between identified barriers and ICF chapters, categorization according to ICF was not pursued further. Instead, new categories were developed based on areas of life participation primarily impaired by specific barriers. For hearing impairments, this resulted in three categories, for visual impairments in five (including those also formed for hearing impairments).

Data Collection for Technology Analysis. In the previous project, a technology database was created [4]. The database (last update cycle December 2023-September 2024) was searched for relevant AI assistance. If the product or project description was explicitly for persons with visual impairment or blindness or persons with hearing impairment or deafness, the AI technology was selected for analyses. Consequently, AI technologies such as Claude, ChatGPT, and Copilot were excluded from the analysis.

Technology Analysis. AI assistance was assessed quantitatively and qualitatively. For all technology assessments, online manufacturer data or project information was used. Market maturity and availability were analyzed to quantify demand coverage. The functions provided to users by AI assistance were identified and described in a more in-depth approach. A function describes the service an AI technology provides for assisting persons with disabilities. AI-supported assistance technologies can have one or more functions. To determine which functions were supported by AI and to describe the functions accordingly, data scientists used the definition and systematization of AI from Russell & Norvig [18]. The distinction between AI functions (‘functions that use methods of […] AI to assist persons with specific disabilities in meeting their work needs, requirements and tasks’) and digital (lacking AI methods) was subject to uncertainty as some functions can be realized with the use of AI methods and without it and not all given information was sufficient. Similar functions using AI methods were grouped

into AI functions. The search included barriers and AI assistance related to workplace learning and vocational rehabilitation. Features not specifically designed for persons with disabilities (e.g., translation, audio recording) were not considered. Describing AI functions and analyzing the number of AI assistance technologies with these functions and their availability in the market was necessary to identify gaps regarding the challenges and research and development needs for AI. For the gap analysis, four tiers were used to describe the technological state of AI functions for both target groups. Those tiers were then compared to the challenges they addressed.

3 Results

3.1 Barriers Encountered by Persons with Visual and Hearing Impairments

Visual Impairment and Blindness. 25 publications along with the results of the Design Thinking workshops (to be referenced as [KI.ASSIST]) were evaluated. Five categories were identified, but only four are related to learning and presented here:

Sensory Perception (Information Intake and Processing). Individuals often face tasks requiring sensory perception and information processing (e.g., from text, images, videos) such as sorting files, filing forms, reading small print [KI.ASSIST], working on a PC or machine. Those relying on screen reader software encounter challenges with collaboratively edited documents, certain elements in common office software or websites [23], charts, visualizations and diagrams [1] and multimedia content, e.g. images containing text [25]. Color perception was stated as another challenge [KI.ASSIST].

Communication as a Sender (Transmission of Information). Respective barriers are twofold: first, the protection of personal data and privacy when interacting with voice assistants [8, 11], and second, challenges associated with speech-to-text conversion, particularly in noisy environments [8] or when users stutter or use dialect [11].

Communication as a Receiver (Reception of Information). The facial recognition of potential discussion partners [KI.ASSIST] and their emotional expressions and non-verbal messages [17] during conversations were named as challenges for blind persons.

Information Technology Use. Integrating assistive functions into existing devices is preferred, as switching to new devices may be challenging [2]. Permanently holding a smartphone for navigation or object recognition is described as uncomfortable due to theft or falling risks [24]. However, the visibility of wearable devices can lead to stigmatization [24, KI.ASSIST]. Also, non-accessible tech support was mentioned [23].

Hearing Impairments and Deafness. Evaluation of 29 studies along with the results of the Design Thinking workshops [KI.ASSIST] resulted in three categories of barriers:

Sensory Perception. Perceiving information is challenging for persons with hearing impairments, because much information in our surroundings is given acoustically, like information at railway stations [6] or explanations in videos. Also, it can be problematic when warning signals or technical malfunctions, are not heard [13]. Non-verbal background noises at the workplace can lead to listener fatigue and limit the ability to hear

relevant noises or voices. For deaf persons who cannot read written language, many websites, work documents are inaccessible [15]. This extends to learning materials.

Communication as a Sender. As communication situations are bidirectional, persons with hearing impairments face barriers when communicating as a sender. Those relying on sign language cannot send messages to many of their coworkers or customers. That is especially the case for telephone calls which do not include videos. In some cases, hearing impairments can lead to speech impairments, thus hindering communication as a sender [KI.ASSIST].

Communication as a Receiver. Persons with hearing impairments are prone to miss more communication opportunities [15]. Telephone calls and missed ringing pose problems [6, 9], because in jobs they are daily tasks. For hard of hearing persons, understanding accents, dialects, or unclear speech and following group conversations with rapidly changing speakers can be challenging [KI.ASSIST], especially in video calls with low-quality microphones [5]. Lip reading "requires constant attention to obtain complete information" and is hard to combine with other tasks [19, p. 1]. In video calls, low-quality cameras or latency issues, low lightning, or speakers not facing the camera, hinder lip reading [12]. Persons using sign language encounter considerable communication barriers at work, as few hearing individuals know sign language and interpreters are often not available [14].

3.2 Matching Barriers and AI Assistance for Both Target Groups

The results for AI assistance for both groups are shown in Table 1, containing for each category the number of technologies, relevant AI functions and maturity assessment. Gap analysis regarding barriers and AI functions resulted in four tiers for both target groups: tier 1 'maturity', tier 2 'development and transfer', tier 3 'research' and tier 4 'no AI assistance found". While many of the barriers for persons with visual impairments are already tackled by AI functions, none were identified for barriers using screen readers for collaboratively edited documents, certain elements in common office software or websites, public transport, data and privacy protection during voice input into voice assistants, and voice input in noisy environments. For persons with hearing impairments, no AI functions were identified to tackle missing communication possibilities (due to not hearing being spoken to), combining lip reading with other tasks and following group conversations (e.g., in video conferences).

4 Discussion and Conclusion

This short paper presents the first results regarding barriers persons with seeing and hearing impairments face in work and vocational rehabilitation, existing AI assistance that can help overcome these, and provides first insights into gaps and possible research and development needs. For persons with visual impairments, further research (tier 3) might be needed in emotion recognition of others and their description to improve social interactions, and in access to visual information (e.g., charts, diagrams). Our analysis showed no AI assistance for using certain elements in office software, on websites,

Table 1. Barriers and AI support for persons with visual and hearing impairments

Cluster	AI assistance for persons with visual impairments (n = 47)					AI assistance for persons with hearing impairments (n = 22)				
	AI Functions	proportion (absolute number) of AT				AI Functions	proportion (absolute number) of AT			
		product	research project	**total**	total per cluster		product	research project	**total**	total per cluster
Communication as a sender	Recognize speech and translate to text	4% (2)	-	**4% (2)**	6% (3)	Recognize German sign language and translate it into text or speech	-	14% (3)	**14% (3)**	18% (4)
	Predict words during writing	2% (1)	-	**2% (1)**		Convert unintelligible pronunciation into intelligible pronunciation	5% (1)	-	**5% (1)**	
Communication as a receiver	Recognize and describe emotions in others	-	6% (3)	**6% (3)**	6% (3)	Recognize speech and translate to text	32% (7)	-	**32% (7)**	82% (18)
						Improve speech intelligibility	14% (3)	14% (3)	**27% (6)**	
						Transcribe phone calls	9% (2)	-	**9% (2)**	
						Focus auditory attention on one person	5% (1)	5% (1)	**9% (2)**	
						Convert spoken language or texts into German sign language	-	9% (2)	**9% (2)**	
						Improve speech intelligibility for phone calls	-	9% (2)	**9% (2)**	
						Automatically subtitle videos	5% (1)	-	**5% (1)**	
Sensory perception	Recognize and read aloud texts	28% (13)	6% (3)	**34% (16)**	55% (26)	Filter out background noise	9% (2)	5% (1)	**14% (3)**	27% (6)
	Recognize and describe objects	17% (8)	6% (3)	**26% (12)**		Adjust individual sound settings for hearables	5% (1)	5% (1)	**9% (2)**	
	Recognize and name familiar faces	6% (3)	6% (3)	**13% (6)**		Detect and alert to dangerous environmental noises	5% (1)	-	**5% (1)**	
	Detect and name colors	9% (4)	2% (1)	**11% (5)**		Recognize spoken language and provide contextual information	-	5% (1)	**5% (1)**	
	Navigate in texts by voice commands	9% (4)	2% (1)	**11% (5)**						
	Model surrounding acoustically	2% (1)	4% (2)	**6% (3)**						
	Recognize and describe scenery	4% (2)	-	**4% (2)**						
	Identify light sources	2% (1)	2% (1)	**4% (2)**						
	Convert graphics into haptic Braille patterns	2% (1)	-	**2% (1)**						
	Recognize and summarize documents	2% (1)	-	**2% (1)**						
	Recognize and count faces	2% (1)	-	**2% (1)**						
	Recognize and read aloud handwriting	-	2% (1)	**2% (1)**						
	Summarize and describe charts, visualizations and diagrams	-	2% (1)	**2% (1)**						
Information technology use	Control devices with voice commands	6% (3)	2% (1)	**9 % (4)**	15% (7)					
	Read out website content	4% (2)	-	**4% (2)**						
	Adapt websites for screen readers	2% (1)	-	**2% (1)**						
	Describe pictures on websites	2% (1)	-	**2% (1)**						
	Interact with objects	-	2% (1)	**2% (1)**						

Product = either available/market-ready or announced; Research projects = ongoing projects and completed projects with prototype available; No information on market readability available for n = 1 AT (*recognize and describe objects*); One technology may contain one or more AI functions and offer support in one or more clusters; Category "mobility" for persons with visual impairments (N = 22 AT) not included in this table due to minor relevance for the focus on work, learning or Vocational Education and Training settings.

and for collaboratively edited documents (tier 4), tools also used for learning. Persons relying on screen readers are often dependent on developers to ensure compatibility. Speech control, in combination with generative AI, could offer significant potential to overcome these barriers. For persons with hearing impairments, research projects (tier 3) developing prototypes for automatic recognition and translation of sign language in

speech or text and vice versa are promising to enhance participation and independence of deaf persons in work, education, and social life. No AI assistance (tier 4) was found to help following group conversations with multiple, fast-changing speakers (e.g., in video calls). Existing functions in apps to identify/ highlight speakers depend on other participants using specific apps and features. AI-based speaker recognition integrated into common video conference tools can enhance independence. Communication as a receiver and obtaining information from the world around them are central challenges for both groups. These might be a good starting point for AI research and development.

Acknowledgments. This study was funded by German Federal Ministry of Labour and Social Affairs (BMAS).

Disclosure of Interests. All authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alam, M., Islam, S., Hoque, E.: SeeChart: enabling accessible visualizations through interactive natural language interface for people with visual impairments. In: Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23), pp. 46–64. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3581641.3584099
2. Ashraf, M., Hasan, N., Lewis, L., Hasan, M., Ray, P.: A systematic literature review of the application of information communication technology for visually impaired people. Int. J. Disabil. Manag. **11**(6), 1–18 (2017). https://doi.org/10.1017/idm.2016.6
3. Beudt, S.: Künstliche Intelligenz und Inklusion in der Arbeitswelt – Leitlinien und Kompetenzen für die KI-gestützte Förderung beruflicher Teilhabe. Digitales Deutschland (2024). https://digid.jff.de/ki-expertisen/inklusion
4. Blanc, B., Beudt, S.: Monitoring KI-gestützter Assistenztechnologien für Menschen mit Behinderungen. Stand der Entwicklungen und Trends. Ergebnisbericht des Projekts KI.ASSIST. Bundesverband Deutscher Berufsförderungswerke e. V. (2022)
5. Dirks, S., Kurth, F.: Working from home in the COVID-19 pandemic - which technological and social factors influence the working conditions and job satisfaction of people with disabilities? In: Miesenberger, K., Kouroupetroglou, G., Mavrou, K., Manduchi, R., Rodriguez, M.C., Penáz, P. (eds.), Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022 July 11–15, 2022, Proceedings, Part II (1st ed. 2022.), pp.183–191. Springer, Lecco (2022)
6. Haynes, S., Linden, M.: Workplace accommodations and unmet needs specific to individuals who are deaf or hard of hearing. Disabil. Rehabil. Assist. Technol. **7**(5), 408–415 (2012). https://doi.org/10.3109/17483107.2012.665977
7. Kähler, M., Feichtenbeiner, R., Beudt, S.: Facilitating the implementation of ai-based assistive technologies for persons with disabilities in vocational rehabilitation: a practical design thinking approach. In: International Conference on Artificial Intelligence in Education, pp. 224–228. Springer, Cham (2021)
8. Komninos, A., Stefanis, V., Garofalakis, J.: A review of design and evaluation practices in mobile text entry for visually impaired and blind persons. Multimodal Technol. Interact. **7**(22) (2023). https://doi.org/10.3390/mti7020022
9. Lange, M., Decker, M.: Mit Hörschädigung im Job: Unterstützung mit Luft nach oben. Spektrum Hören **40**(5), 32–35 (2020)

10. Long, D., Magerko, B.: What is AI literacy? competencies and design considerations. In: Bernhaupt, R., et al. (eds.), Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16. ACM (2020). https://doi.org/10.1145/3313831.3376727
11. May, S., Szczypien, N., Vahldiek, K., Klawonn, F.: Die Rolle Künstlicher Intelligenz in Assistiven Technologien. In: Assistive Technologien im Sozial- und Gesundheitssektor, pp. 51–77. Springer Fachmedien, Wiesbaden (2022)
12. McNamara, K.A., Stanch, P.M.: Accommodating workers with disabilities in the post-Covid world. J. Occup. Environ. Hyg. **18**(4–5), 149–153 (2021). https://doi.org/10.1080/15459624.2021.1902531
13. Morata, T.C., Themann, C.L., Randolph, R.F., Verbsky, B.L., Byrne, D.C., Reeves, E.R.: Working in noise with a hearing loss: perceptions from workers, supervisors, and hearing conservation program managers. Ear Hear. **26**(6), 529–545 (2005). https://doi.org/10.1097/01.aud.0000188148.97046.b8
14. Murray, J., Klinger, L., McKinnon, C.: The deaf: an exploration of their participation in community life. OTJR: Occup. Participat. Health **27**(3), 113–120 (2007). https://doi.org/10.1177/153944920702700305
15. Perkins-Dock, R.E., Battle, T.R., Edgerton, J.M., McNeill, J.N.: A survey of barriers to employment for individuals who are deaf. JADARA **49**(2), 3 (2015)
16. Pinkwart, N., Beudt, S.: Künstliche Intelligenz als unterstützende Lerntechnologie. Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO (2020). http://publica.fraunhofer.de/dokumente/N-624584.html. Accessed 04 Apr 2025
17. Pushpalatha, M.N., Meherishi, H., Vaishnav, A., Anurag Pillai, R., Gupta, A.: Facial emotion recognition and encoding application for the visually impaired. Neural Comput. Appl. **35**(1), 749–755 (2023)
18. Russell, S. J., Norvig, P.: Artificial Intelligence: A Modern Approach. 3rd edn. Prentice Hall, Upper Saddle River (2010)
19. Santoso, Y.B., Astuti, E.Y., Ratnawulan, T., Khoeriah, N.D., Hakim, L.L.: Development assistive technology for students with hearing impairments. J. Phys: Conf. Ser. **1539**(1), 12042 (2020)
20. Scanlan, M.: Reassessing the disability divide: unequal access as the world is pushed online. Univ. Access Inf. Soc. **21**(3), 725–735 (2022). https://doi.org/10.1007/s10209-021-00803-5
21. Thieke-Beneke, M., Stock, J., Lippa, B., Biedermann, J., Stähler, L., Feichtenbeiner, R.: Die KI.ASSIST Lern- und Experimentierräume zur Erprobung KI-gestützter Assistenztechnologien. Von der Konzeption bis zur Umsetzung. Ergebnisbericht des Projekts KI.ASSIST. Bundesverband Deutscher Berufsförderungswerke e.V. (2022)
22. United Nations Department of Economic and Social Affairs. The Sustainable Development Goals Report 2023: Special Edition. United Nations. https://unstats.un.org/sdgs/report/2023/. Accessed 04 Apr 2025
23. Wahidin, H., Waycott, J., Baker, S.: The challenges in adopting assistive technologies in the workplace for people with visual impairments. In: Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI '18), pp. 432–442. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3292147.3292175
24. Walle, H., De Runz, C., Serres, B., Venturini, G.: A survey on recent advances in AI and vision-based methods for helping and guiding visually impaired people. Appl. Sci. **12**(5), 2308 (2022). https://doi.org/10.3390/app12052308
25. Whitney, G., Kolar, I.: Am I missing something? Experiences of using social media by blind and partially sighted users. Univ. Access Inf. Soc. **19**, 461–469 (2019)
26. Whittaker, M., et al.: Disability, bias, and AI. 6 Now Inst. **8**, 11 (2019)

AI-Powered Math Learning: Evaluating the Impact of a Personalized Tutoring Platform that Responds to Affect

Mingyu Feng[1](✉), Natalie Brezack[1], Chunwei Huang[1], Megan Schneider[1], Ivon Arroyo[2], Beverly Woolf[2], and Danielle Allessio[2]

[1] WestEd, San Francisco, CA 94107, USA
{mfeng,nbrezac,chuang,mschnei}@wested.org
[2] University of Massachusetts Amherst, Amherst, MA 01003, USA
{ivon,bev}@cs.umass.edu, allessio@umass.edu

Abstract. Math education remains a critical focus for national education improvement in the U.S. Given the benefits of personalized learning and affective interventions for math learning, schools and districts seek tools with such support to engage students. We report on recent findings from a large-scale randomized controlled trial of a technology-based learning platform that leverages AI-based adaptive technology to offer personalized content, remedial tutoring, immediate feedback, and affective support for students and provides learning analytics on student performance and effort for teachers. The experimental study involved 64 math teachers. Analyses suggest that implementing the program helped teachers adjust their practices, and students who used the program more performed better than the comparison group after controlling for prior performance.

Keywords: Middle school · Math learning · Personalized learning · Engagement

1 Introduction

An urgent need exists to implement innovative approaches to improve math performance. The U.S. National Assessment of Educational Progress (NAEP) showed that only 28% of eighth grade students performed at or above proficiency in math in 2024 [1]. Student math achievement is crucial in middle school; if students fall behind, they are less likely to succeed in algebra and subsequent math classes [2]. Notable losses in student motivation, engagement, and achievement occur in middle school [3], and students' subjective value and math self-concept decline [4]. Student affect and engagement need to be addressed as a pathway to improve math learning.

Research has demonstrated the potential of intelligent educational technologies to close achievement gaps and transform teacher practices (e.g., [5–8]). MathSpring uniquely incorporates affective support during math problem-solving and enables teachers to view rich data on student effort and emotions. Few math platforms integrate support for building positive affect into math (e.g., teachers can send students encouraging

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 118–125, 2025.
https://doi.org/10.1007/978-3-031-99264-3_15

messages in IXL; students have a digital companion in ST Math). Relatedly, research is limited regarding how customized affective interventions impact student learning in middle school. To further this discourse, we conducted a multi-year, large-scale randomized controlled trial (RCT) of the MathSpring program from 2020–2023.

MathSpring is an intelligent tutoring system that provides students with personalized content, remedial tutoring, and metacognitive and affective support (Fig. 1). An educator dashboard provides learning analytics reports on students' performance and effort for teachers to monitor individual and class progress and to assign problems for additional reinforcement. It incorporates metacognitive and self-regulation theories by providing students with tools that encourage self-reflection (e.g., effort charts, feedback, learning companion prompts). These tools actively monitor and regulate students' learning process, fostering a strategic and effective approach to math problem-solving.

The program's Theory of Change includes specific supports for students at risk of falling behind or who have negative attitudes towards math (Fig. 1; [9]). MathSpring personalizes students' learning experience utilizing measures of within-system student learning behaviors to tailor feedback and responses to each student's needs. As a supplemental program, MathSpring includes problems that address grade-level math topics in the U.S. Common Core State Standards for Mathematics. The program has demonstrated evidence in prior small-scale studies [10, 11].

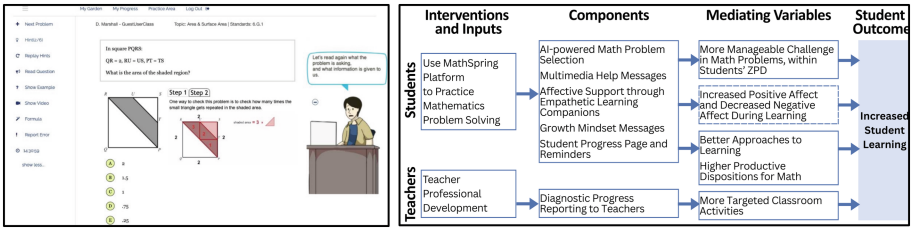


Fig. 1. Left: MathSpring learning environment student view, with examples and video hints available in the left menu, hints visible in the middle, and a learning companion avatar. Right: MathSpring Theory of Change.

In the RCT, we investigated the extent to which the program supports math learning and the nature of teachers' implementation. This paper reports on the findings from recently completed analyses and addresses the following research questions (RQs):

- RQ1: What is the impact of MathSpring on student math learning outcomes after one year of usage?
- RQ2: Does MathSpring influence teachers' instructional practices in supporting students' engagement and effort in math problem solving?
- RQ3: What is the relationship between the amount of MathSpring usage and students' learning outcomes?

2 Role of AI and Affect in Online Learning

MathSpring is a web-based intelligent tutor [10–12]. Students receive math problems tailored to their performance and effort with immediate feedback, learning supports (hints, worked examples, tutorial videos), and socioemotional support. Animated characters called learning companions (LCs; Fig. 2) use gestures to visually reflect students' most recent emotion, offer advice and encouraging messages, and are treated as "social beings" [11–14]. LCs express meta-cognitive and affective feedback and talk to students about the importance of perseverance, effort, and growth mindset [15]. Students are encouraged to review their progress, where their effort is visualized via a plant that blooms to promote self-reflection and responsibility over their learning (Fig. 2).

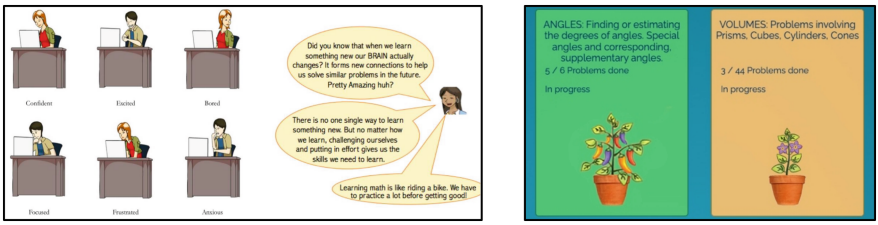


Fig. 2. Left: LCs in MathSpring show a range of emotion and growth mindset messages. Right: Portion of student progress page.

MathSpring provides detailed student- and class-level analytics, including estimates of student learning overall, progress per skill over time, and effort data to inform teachers' adjustments to instruction. Teachers can view data to determine whether students are putting in effort. Teachers can input a class goal for the number of problems solved and display this goal to the class. Using a bottom-up selection process, teachers chose math content (topics and problems) based on grade-level appropriateness. Throughout the year, teachers revised the content; this responsive approach to curriculum development supported adjustments based on student progress and learning needs. At the end of the school year, many teachers activated all topics to provide comprehensive review.

On the backend, students are continuously assessed at the click level as the system gathers log data to maintain students within their zone of proximal development [16]. MathSpring maps student skills required at each grade level onto a fine-grained cognitive model of 214 math topics and automatically assesses students' knowledge based on the model and their level of mastery. Mastery is calculated using Bayesian estimation and updated based upon students' performance on each problem in a content area, accounting for student engagement [17]. An effort-based tutoring algorithm considers a student's recent performance on each topic and level of effort exerted (e.g., skipping, not reading, solving with hints) to optimize problem selection within content areas [18].

MathSpring was built upon theoretical foundations and empirical research. *Metacognitive and self-regulative* behaviors influence students' academic success [19]. Several models [20, 21] define the learning process as a flow model where students use strategies to produce an outcome that is subjected to external feedback. *Academic mindset* is defined as ability beliefs combining with subjective task value to establish expectations

of success [22]. Human qualities, such as intellectual skills, can be cultivated through *effort* [15], and students' *mindset* impacts their help-seeking behavior. When help is abused or avoided, it might become detrimental to learning [23]. *Engagement, motivation, and interest* are precursors to learning, effortful problem solving, and deep thinking [24]. *Emotions* positively or negatively impact student learning and achievement [25] and can be influenced by the presence of someone who cares, which can encourage persistence [26]. *Mastery learning* improves achievement than more traditional teaching [27], and *feedback* improves learning (e.g., [28, 29]).

3 Research Design

The RCT employed a multisite, clustered randomized experimental design: teachers were randomly assigned to either the treatment or the business-as-usual control condition. We compiled publicly available district, school, demographic, and prior achievement data, then used a hierarchical cluster analysis to group teachers into strata. Teachers within each stratum were randomly assigned to treatment or control conditions. Teachers in the treatment group were expected to use MathSpring at least once every other week for one school year (October through May). The selection and assignment of topics was teacher-driven, based on grade-level appropriateness for their specific classes. In the control condition, teachers used their existing instructional practices and supplemental technologies (other than MathSpring).

Sixty-four fifth- or sixth-grade math teachers from one U.S. state (34 treatment, 30 control) enrolled in one of three cohorts. Teachers came from 47 schools in 41 districts with diverse backgrounds. In participating schools, 69% of students were White, 7% were Black, and 15% were Hispanic. 56% of students performed below proficiency level. The development team provided professional development (PD) for treatment teachers prior to implementing MathSpring. PD began with a one-hour session over videoconference. Teachers then worked through PD modules asynchronously, following a monthly *learn*, *do*, and *reflect* structure: *learn:* complete online module lessons and activities; *do:* interact with the teacher tools; *reflect:* reflect on what they did.

4 Data, Analysis, and Results

4.1 Data Collection

Student Learning Outcome. The primary measure of student math achievement was the mandated statewide standardized assessment: Massachusetts Comprehensive Assessment System (MCAS) [30]. The computerized math test was given to all students in grades 3–8 and 10 in May at the end of the school year and included a variety of question formats, including multiple choice, short answer, and essays. We collected MCAS data for a total of 1,677 students (treatment n = 788; control n = 889). Students' performance on the Online Mathematics Diagnostic Testing Project Math Readiness Test (MRT) [31], a 40-question diagnostic assessment administered in September at the beginning of the school year, was used as a baseline math achievement measure.

Implementation Measures. We conducted semi-structured interviews with a representative sample of 35 teachers across districts and schools with varied MathSpring usage levels to understand implementation and its impact on teacher practices.

Student and Teacher Use of MathSpring. MathSpring logged electronic use records, available for 821 treatment students and 27 treatment teachers. Student data included login time, session duration, number of problems attempted and answered correctly and incorrectly, and student-to-system interactions (e.g., requesting hints). Teacher use data included the dates each math topic was made available, math topic assigned, and reports opened. These data allowed us to estimate the dosage of student exposure and whether the program was used following the implementation model.

4.2 Analysis and Results

To evaluate MathSpring's efficacy for improving math learning (RQ1), we conducted an intent-to-treat analysis using two-level hierarchical linear regression models (HLM) to account for data clustering (students within teachers). We modeled the mean differences in the outcome measures between students in the treatment and control groups, controlling for teacher-level characteristics, pretest MRT scores, and other student- and teacher-level covariates to improve precision. Overall, students in the treatment group outperformed those in the control group on the MCAS, accounting for baseline MRT. The difference was not significant at the 0.05 level.

Qualitative and descriptive analyses were conducted on interview transcripts to address RQ2 (teacher practices). Researchers coded interviews according to four socio-emotional learning-related sub-categories (Table 1). Teachers in the treatment group mentioned each sub-category significantly more often than those in the control group [32]. Using MathSpring encouraged teachers to incorporate socio-emotional learning topics and strengthened their ability to deliver this type of instruction with their students. Teachers reported using MathSpring's effort data to spark class discussions around effort, which can help students develop self-regulated learning strategies.

Table 1. The number (and percentage) of instances in which each of the socio-emotional sub-categories were mentioned by teachers [32].

SEL sub-categories	Treatment	Control
The importance of effort	22 (66.7%)	7 (29.2%)
Productive problem-solving strategies	25 (75.8%)	10 (41.7%)
Setting goals	15 (51.5%)	1 (4.2%)
Emotions	21 (63.6%)	11 (45.8%)

To address RQ3 (impact of usage on learning), we used the total amount of time students engaged in MathSpring as the indicator of usage and split the treatment teachers into three subgroups based on the average of their students' usage: high (above 150

min), medium (100–150 min), and low (less than 100 min). Subgroups were determined through discussions with the developer about sufficient implementation time; students averaged 190 min (maximum 978 min) of total use. This low usage of MathSpring was due, in part, to factors such as district policy or the introduction of competing programs during the COVID pandemic. We conducted an exploratory analysis to compare student learning outcomes of each subgroup against the control group, using the two-level HLM models, controlling for the same covariates as in RQ1. The results indicated that students in the high-use subgroup performed marginally better that the students in the control group, with a meaningful effect size ($p = 0.059$; effect size $= 0.27$). The differences between the medium or low use subgroups and the control group were not statistically significant, although students in both treatment subgroups outperformed those in the control group.

5 Conclusion

Visions abound for leveraging educational technology to enhance math outcomes, aiding students in overcoming setbacks caused by the pandemic and reengaging students in math learning. MathSpring illustrates how effective design strategies and learning theories can be incorporated to empower technology-based programs to support math learning, emotions, and effort. MathSpring shows potential for improving learning when implemented in authentic school settings at scale; students who used MathSpring demonstrated stronger performance on the state standardized test, although the evidence was not conclusive. Yet, exploratory analysis suggests that students with more exposure to MathSpring demonstrated marginally better math learning than those in the control group, scoring 0.27 standard deviations higher on the end-of-year state test.

Of note, qualitative analyses indicated that using MathSpring influenced teaching practices: Teachers who used MathSpring incorporated emotion and effort more frequently in discussions with their students than did teachers in the control group. Teachers reported that MathSpring changed their classroom teaching behavior and encouraged them to discuss topics such as goal setting, effort, and productive problem-solving strategies. This represents a significant advancement beyond traditional intelligent tutoring systems that typically prioritize content delivery and personalization without addressing the equally important social-emotional dimensions of educational experience.

The reported results are bound by limitations related to the disruption of classroom instruction due to the pandemic, which presented significant challenges for recruitment, data collection, and participant retention. Overall, the dosage of student exposure to MathSpring was relatively low, which may explain the null main effect. Comparable studies with significant results reported 5–10 h of student exposure to the intervention; this effect from 2.5 h shows the program's potential. Further analysis is ongoing to better understand the nature of use during and after the pandemic.

Acknowledgements. This material is based on work supported by the Institute of Education Sciences of the U.S. Department of Education under Grant 305A190256. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the funders.

Disclosure of Interests. *MathSpring* is a free service provided by University of Massachusetts Amherst, Amherst. The study was conducted independently by WestEd. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. National Assessment of Educational Progress (NAEP): 2024 mathematics grade 8 assessment results (2024)
2. Siegler, R.S., et al.: Early predictors of high school mathematics achievement. Psychol. Sci. **23**, 691–697 (2012)
3. Eccles, J.S., Midgley, C.: Changes in academic motivation and self-perceptions during early adolescence. In: Montemayor, R., Adams, G.R., Gullotta, T.P. (eds.) Advances in Adolescent Development: From Childhood to Adolescence, pp. 134–155. Sage, Newbury Park (1990)
4. Royer, J.M., Walles, R.: Influences of gender, motivation and socioeconomic status on mathematics performance. In: Berch, D.B., Mazzocco, M. (eds.) Why is Math so Hard for Some Children, pp. 349–368. Paul H. Brookes Publishing Co, Baltimore (2007)
5. Roschelle, J., et al.: Integration of technology, curriculum, and professional development for advancing middle school mathematics: three large-scale studies. Am. Educ. Res. J. **47**(4), 833–878 (2010)
6. Roschelle, J., Feng, M., Murphy, R., Mason, C.: Online mathematics homework increases student achievement. AERA Open J. **2**(4), 1–12 (2016)
7. Feng, M., Brezack, N., Huang, C., Collins, K.: Long-term effects of an online math tool on U.S. adolescents' achievement. Br. J. Educ. Technol. (2025)
8. Woolf, B., et al.: FACE READERS: the frontier of computer vision and math learning. Comput. Sci. **6** (2023). https://commons.clarku.edu/faculty_computer_sciences/6
9. Woolf, B., et al.: The effect of motivational learning companions on low achieving students and students with disabilities. In: Proceeding of the 10th International Conference on Intelligent Tutoring Systems, pp. 327–337. ITS, Pittsburgh, PA (2010)
10. Arroyo, I., et al.: Repairing disengagement with non-invasive intervention. In: Proceedings of the 13th International Conference on Artificial Intelligence in Education, pp. 195–202. AIED, Amsterdam, Netherlands (2007)
11. Arroyo, I., Burleson, W., Tai, M., Muldner, K., Woolf, B.P.: Gender differences in the use and benefit of advanced learning technologies for mathematics. J. Educ. Psychol. **105**(4), 957–969 (2013)
12. Gupta, A., et al.: Affective teacher tools: affective class report card and dashboard. In: Proceedings of the 22nd International Conference on Artificial Intelligence in Education, pp. 178–189. AIED, Utrecht, The Netherlands (2021)
13. Lee, W., et al.: Measurements and interventions to improve student engagement through facial expression recognition. In: Sottilare, R.A., Schwarz, J. (eds.) Adaptive Instructional Systems, pp. 286–301. Springer, Cham (2022)
14. Delgado, K., et al.: Student engagement dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3628–3636 (2021)
15. Dweck, C.S.: Mindset. Random House, New York (2006)
16. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge (1978)
17. Arroyo, I., Woolf, B.P., Burelson, W., Muldner, K., Rai, D., Tai, M.: A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. Int. J. Artif. Intell. Educ. **24**(4), 387–426 (2014)

18. Arroyo, I., Mehranian, H., Woolf, B.P.: Effort-based tutoring: an empirical approach to intelligent tutoring. In: The 3rd International Conference on Educational Data Mining. Educational Data Mining, Pittsburgh, PA, USA (2010)
19. Dignath, C., Buettner, G., Langfeldt, H.: How can primary school students learn self-regulated learning strategies most effectively?: a meta-analysis on self-regulation training programmes. Educ. Res. Rev. **3**(2), 101–112 (2008)
20. Butler, D.L., Winne, P.H.: Feedback and self-regulated learning: a theoretical synthesis. Rev. Educ. Res. **65**(3), 245–281 (1995)
21. Zimmerman, B.J., Moylan, A.R.: Self-regulation: where motivation and metacognition intersect. In: Hacker, D.J., Dunlosky, J., Arthur C.G. (eds.) Handbook of Metacognition in Education, pp. 311–328. Routledge (2009)
22. Wigfield, A., Eccles, J.S.: Expectancy–value theory of achievement motivation. Contemp. Educ. Psychol. **25**(1), 68–81 (2000)
23. Aleven, V., McLaren, B., Roll, I., Koedinger, K.: Toward meta-cognitive tutoring: a model of help-seeking with a cognitive tutor. Int. J. Artif. Intell. Educ. (IJAIED) **16**(2), 101–128 (2006)
24. Rose, D.H., Meyer, A.: Teaching every student in the Digital Age: Universal design for learning. Assoc. for Supervision & Curriculum Development, Alexandria, VA, USA (2002)
25. Pekrun, R., Goetz, T., Daniels, L., Stupinsky, R., Perry, R.: Boredom in achievement settings: exploring control–value antecedents and performance outcomes of a neglected emotion. J. Educ. Psychol. **102**(3), 531–549 (2010)
26. Burleson, W., Picard, R.W.: Evidence for gender specific approaches to the development of emotionally intelligent learning companions. IEEE Intell. Syst. Spec. issue Intell. Educ. Syst. **22**(4), 62–69 (2007)
27. Koedinger, K., Aleven, V.: Exploring the assistance dilemma in experiments with cognitive tutors. Educ. Psychol. Rev. **19**, 239–264 (2007)
28. Mendicino, M., Razzaq, L., Heffernan, N.T.: A comparison of traditional homework to computer supported homework. J. Res. Technol. Educ. (JRTE) **41**(3), 331–358 (2009)
29. Fyfe, E.R., Rittle-Johnson, B.: The benefits of computer-generated feedback for mathematics problem solving. J. Exp. Child Psychol. **147**, 140–151 (2016)
30. Massachusetts Department of Elementary and Secondary Education (MDESE): 2023 Next-Generation MCAS and MCAS-Alt Technical Report (2024). https://www.doe.mass.edu/mcas/tech/2023-nextgen-tech-report.pdf. Accessed 1 Apr 2025
31. Anthony, M.: Consequential-related validity evidence for MDTP tests. Data submitted in response to California Community Colleges Assessment Standards for renewal of placement test instruments (2005). http://mdtp.ucsd.edu
32. Brezack, N., Chan, W., Feng, M.: Student effort and progress learning analytics data inform teachers' SEL discussions in math class. In: 14th Proceedings of the Learning Analytics and Knowledge Conference, pp. 338–348. Association for Computing Machinery, New York (2024)

Tell Me Who Your Students Are: GPT Can Generate Valid Multiple-Choice Questions When Students' (Mis)Understanding is Hinted

Machi Shimmei[1,5(✉)], Masaki Uto[2], Yuichiroh Matsubayashi[1,5], Kentaro Inui[1,3,5], Aditi Mallavarapu[4], and Noboru Matsuda[4]

[1] Tohoku University, Sendai, Japan
machi.shimmei.e6@tohoku.ac.jp
[2] The University of Electro-Communications, Tokyo, Japan
[3] MBZUAI, Abu Dhabi, United Arab Emirates
[4] North Carolina State University, Raleigh, USA
[5] RIKEN, Wako, Japan

Abstract. The primary goal of this study is to develop and evaluate an innovative prompting technique, AnaQuest, for generating multiple-choice questions (MCQs) using a pre-trained large language model. In AnaQuest, the choice items are sentence-level assertions about complex concepts. The technique integrates formative and summative assessments. In the formative phase, students answer open-ended questions for target concepts in free text. For summative assessment, AnaQuest analyzes these responses to generate both correct and incorrect assertions. To evaluate the validity of the generated MCQs, Item Response Theory (IRT) was applied to compare item characteristics between MCQs generated by AnaQuest, a baseline ChatGPT prompt, and human-crafted items. An empirical study found that expert instructors rated MCQs generated by both AI models to be as valid as those created by human instructors. However, IRT-based analysis revealed that AnaQuest-generated questions—particularly those with incorrect assertions (foils)—more closely resembled human-crafted items in terms of difficulty and discrimination than those produced by ChatGPT.

Keywords: Educational question generation · analytical multiple-choice questions · large language model · prompt engineering · item response theory

1 Introduction

Multiple-choice questions (MCQ) are among the most widely used types in authentic educational settings [1–3]. While MCQs take many forms, this study focuses on *analytical MCQs* on a complex topic with a stem: "Which of the following are correct about [topic]?". Their answer choices consist of *assertions* that require analytical reasoning rather than mere fact recall to be answered correctly. The effectiveness of analytical MCQs depends on the "validity" of answer choices. Validity can be viewed from multiple perspectives, one of which is ensuring that the choices reflect students' understanding

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 126–134, 2025.
https://doi.org/10.1007/978-3-031-99264-3_16

and misconceptions of the topics being assessed. Foils (or "distractors") should strike a balance by being challenging enough to address underlying misconceptions without being too obvious or excessively difficult.

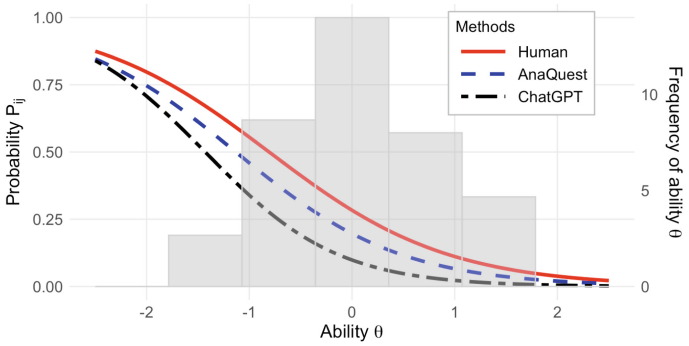


Fig. 1. Item characteristic curve of foils generated by different sources (Human, AnaQuest, and ChatGPT). The mean of discrimination and difficulty params estimated by a two-parameter Item-Response Theory model are used. The shaded histogram shows the sample distribution of the estimated student ability parameter. See details in Sect. 5.1.

Generating foils that tap the ability for analytical reasoning requires knowledge about misconceptions that students might develop. We hypothesize that experienced instructors can create valid answer choices based on their observations about how students struggle with learning complex concepts. Nonetheless, creating MCQs is extremely labor-intensive [4, 5], and research has increasingly explored the generative AI, i.e., pre-trained Large Language Models (LLM) to automate it [6, 7]. Accordingly, a critical question arises: Can LLMs make MCQs with validity comparable to those created by experienced instructors?

Figure 1 illustrates how the probability of selecting a foil varies with the student's latent ability. *The plot reveals a gap in the validity of the foils generated by the bare-bones ChatGPT (vanilla adoption of a pre-trained LLM) and those created by an expert instructor, with the latter exhibiting more desirable foil characteristics.* We hypothesize that the ChatGPT's shortfall stems from its pre-training data, which might lack knowledge of students' misconceptions about the topics being assessed, or from the model's failure to retrieve such knowledge during generation.

To address this issue, we propose AnaQuest, a prompting technique for pre-trained LLMs that incorporates students' understanding of a topic and a learning objective into the prompt. Students' understanding is captured through free-text responses to an open-ended question related to the topic, collected as formative assessments. Figure 1 shows that AnaQuest can generate incorrect assertions that are significantly closer to those created by an instructor than those produced by the bare-bones ChatGPT.

Expert surveys evaluating the pedagogical value of AI-generated questions are a commonly used technique [8, 9]. However, our current study demonstrated that such survey-based evaluations might overlook latent psychometric properties, such as discrimination and difficulty, that distinguish AI-generated from human-crafted questions.

In this study, we propose to assess the validity of AI-generated MCQs by leveraging proximity matrices (e.g., statistical distance), in combination with Item Response Theory, to evaluate how closely their discrimination and difficulty resemble those of MCQs created by experienced instructors—both at the question and foil levels.

The contributions of this study are as follows: (1) We propose AnaQuest, a prompting technique for generating analytical MCQs based on students' understanding of specific learning objectives. To our knowledge, *this is the first approach to use knowledge from formative assessment to generate valid summative assessments.* (2) We evaluated the validity of AI-generated questions and foils by measuring their proximity to human-crafted questions using the Item-Response Theory (IRT) technique on authentic exam data. To our knowledge, *this is the first study to assess the validity of AI-generated MCQs—including both questions and foils—using IRT as a* ***proximity measure****.*

2 Related Work

There has been growing interest in applications of pre-trained LLMs for educational question generation [10–13]. Generating MCQs is one of the most intensively studied topics where how to generate foils (aka distractors) is an essential problem [14, 15]. Researchers have employed linguistic perturbations of correct choices [16–18] and ontological knowledge to generate foils [19–21]. While those techniques are mostly for term-level distractors, we focus on sentence-level assertions.

Van Campenhout *et al.* [22] evaluated the difficulty, engagement, and persistence of AI-generated questions using student response data. Other studies used IRT to estimate question difficulty and discrimination[23, 24]. However, because IRT parameters are not absolute metrics, drawing objective interpretations can be challenging. Our proposed use of IRT-based proximity matrices aims to address this limitation.

3 Analytical Multiple-Choice Question

A multiple-choice question (MCQ) consists of a *stem* and a set of correct and incorrect *choice items.* Incorrect choice items are called distractors or foils. In this paper, we examine a type of MCQs that have sentences as choice items. We particularly focus on MCQs with the following form of stem: "Which of the following are correct about [topic]?" followed by three assertions: A, B, and C, as shown in Fig. 2.

A unique aspect of the current study is that assertions are carefully crafted for a specific learning objective at a higher level of Bloom's Taxonomy. The example shown in Fig. 2 was made for the following learning objective: "Describe the overall structure of contextual design." If different learning objectives are given, different sets of assertions will be crafted for the same topic (i.e., the question stem). The [topic] is a specific theme to be tested. For example, "the overall structure of contextual design" is one ***topic***, and "the meaning of context in the contextual design" is another, though both are about the same ***theme***: "contextual design."

To answer those MCQs correctly, students must analyze and evaluate given information to carry out logical reasoning beyond merely remembering facts. *We hypothesize that when assertions are carefully crafted for learning objectives at a higher level of*

Q:	Which of the following are correct about the overall structure of contextual design?
A:	Contextual design emphasizes understanding user needs in their real environments by interviewing users, modeling their interactions, and iterating design solutions.
B:	The contextual design is participatory method where the primary goal is to solicit ideas about a good interface from users by asking about the deficit of the current design and what users would like to see instead.
C:	The main focus of contextual design is to develop affinity diagrams, and when the affinity diagram is well developed, other models like cultural or sequence-flow models will not be needed.

Fig. 2. Example of an exam question.

Bloom's Taxonomy, answering MCQ requires analytical reasoning skills. We therefore call this type of MCQs the *analytical multiple-choice questions.*

4 Research Questions and Methods

We investigate the following research questions: (RQ1) How do AI-generated questions hold validity relative to those created by an expert instructor? (RQ2) How do expert instructors perceive the validity of AI-generated questions? To address these questions, we conducted an empirical evaluation in a college-level course, comparing MCQs generated by an instructor, AnaQuest, and another baseline generative AI.

4.1 AnaQuest: System Implementation

AnaQuest was implemented using OpenAI GPT-4o with few-shot learning [25] with five examples. It was tasked with generating three correct and incorrect assertions for a given topic and a learning objective. In addition, AnaQuest received students' responses to an open-ended question related to the topic. The model was prompted to ensure that the generated assertions could effectively assess whether a student had achieved the learning objective. It was also instructed to generate incorrect assertions that reflect common student misunderstandings identified in the provided responses.

4.2 Overview of the Evaluation Study

The evaluation study was conducted using the final exam at a college-level course on Human-Computer Interaction (HCI) in fall 2024. Prior to the exam, students answered two open-ended questions per week for eight weeks—i.e., 16 topics were probed with 50 to 56 student responses per topic. As a result, AnaQuest generated 16 multiple-choice questions (MCQs), each with three correct and three incorrect assertions.

OpenAI ChatGPT (4o Dec. 2024 ver.) [26] was used as a baseline generative AI. 16 questions on the same topics as AnaQuest were created using a straightforward prompt:

```
"I need to create a final exam for a graduate level
HCI course. Generate three correct and three incorrect
```

```
choices, each about 30 words, for a multiple choice ques-
tion asking "Which of the following are correct about
[TOPIC]?"".
```

An experienced instructor who has been teaching the HCI course for more than 10 years created eight MCQs with slightly modified topics to avoid duplication with the AI-generated questions, while the overall themes to be tested remained the same. The instructor also reviewed AI-generated questions and selected three assertions per question to be used for the exam while ensuring their adequacy as exam questions.

Two exam versions were created, each containing all eight human-crafted questions, but a different set of five ChatGPT- and five AnaQuest-generated questions, with topic coverage counterbalanced. Each question had 9 answer choices, including all combinations of correct assertions and an "I do not know" option. Students gained 1 point for correct answers, lost 1/7 for incorrect ones, and earned 0 for "I do not know." Students were informed that this scoring policy was used to discourage random guessing.

4.3 Analysis

Multiple-choice questions from three *sources* (AnaQuest, ChatGPT, and Human) were comparatively analyzed using an IRT-based method and a survey study.

The IRT-based method was applied at both the *question-* and *foil-levels*. For question-level analysis, we used the two-parameter logistic (2PL) model [27, 28]. Questions with difficulty parameters beyond mean ability ± 3SD were excluded as outliers. For foil-level analysis, *incorrect assertions* were treated as items. Foil validity was defined as the likelihood that students believe a foil is correct, modeled as a function of their latent ability. Accordingly, we applied the 2PL model to a response matrix indicating which students considered which foils correct. Discrimination and difficulty parameters for each foil were computed using latent ability scores from the question-level analysis. Valid foils are expected to have negative values for both discrimination and difficulty.

The survey study was conducted with three expert instructors who regularly teach the HCI course. Participants were asked to rate each exam question based on the following five criteria, using a 5-Likert scale. The source of question (AnaQuest, ChatGPT, or Human) was hidden in the survey: **S1** (*Answerability*): The question is clear enough to answer. **S2** (*Correctness*): The given answers are correct. **S3** (*Adequate Complexity*): The given answers are adequate, i.e., not too easy, not too difficult, not too confusing, etc. **S4** (*Alignment*): The question is suitable for the learning objective. **S5** (*Adoptability*): I would adopt the question to my course.

5 Results

5.1 Results from the IRT-Based Methods

57 students took the exam: 29 for Ver.A and 28 for Ver.B. According to the overall correctness of student answers, the two versions were equivalent in their complexity; the ratio of correct responses to the total number of responses was 0.56 ± 0.02 and 0.55 ± 0.02 for version A and B, respectively: $t(55) = 0.17$, $p = 0.86$.

Question-Level Item Analysis. Following the rule of outliers (Sect. 4.3), 9 questions (2 Human, 4 ChatGPT, and 3 AnaQuest) were excluded from the analysis. Analysis of discrimination showed that AnaQuest-generated questions had discrimination much more like human-crafted questions than ChatGPT-generated questions; the overall average of discrimination across questions were 0.89 ± 0.25 for AnaQuest vs. 0.85 ± 0.37 for Human vs. 1.07 ± 0.25 for ChatGPT.

To quantitatively compare the proximity of estimated discrimination α and difficulty β among three sources, we estimated the bivariate distributions (α, β). . We then used KL-Divergence (KLD) with Gaussian kernel density estimation [29]. The results showed that KLD between human-crafted and AnaQuest-generated questions (820.96) is notably lower than that between human-crafted and ChatGPT-generated questions (1820.01). KLD between AnaQuest- and ChatGPT-generated questions was 1423.45. *This finding implies that AnaQuest can generate questions with more similar IRT characteristics to human-crafted ones than ChatGPT.*

Foil-Level Item Analysis. Figure 1 shows the item characteristic curve (ICC) to visualize the validity of foils. ICC shows the probability that a student j considers a foil i as a correct assertion as a function of the student's ability θ_j. In Fig. 1, ICC is overlaid on the distribution of student ability, which is shown as a shaded histogram. The plot shows that AnaQuest-generated foils are more closely approximated human-crafted ones than ChatGPT-generated foils. ChatGPT-generated foils exhibited a notably steeper slope (higher discrimination) and much more left-shifted (lower difficulty). *This suggests that ChatGPT-generated foils were more obviously incorrect and, therefore, more likely to be selected only by very low-ability students.*

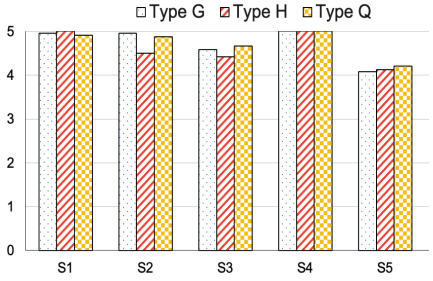


Fig. 3. The average ratings per source—Human (H), AnaQuest (Q), and ChatGPT (G)—across participants for each metric.

To quantify how closely AnaQuest-generated foils resembled human-crafted ones, we again applied the proximity analysis as done in the question-level item analysis. The results showed that the AnaQuest-generated foils had a closer bivariate distribution to Human-crafted ones than ChatGPT; KLD between Human and AnaQuest was 13.66, whereas KLD between Human and ChatGPT was 36.60.

In sum, the foil-level analysis demonstrated that *AnaQuest can produce well-founded foils, fulfilling the critical property of foils for analytical MCQ: low-ability students are more likely to select foils compared to high-ability students.* Furthermore, the analysis revealed that *foils generated by AnaQuest exhibited an appropriate power of discrimination and the level of difficulty, more closely resembling those of human-crafted foils than ChatGPT*.

5.2 Results from Instructor Survey

Figure 3 shows the average ratings by source (Human, AnaQuest, and ChatGPT) across three survey participants, based on the five criteria mentioned in Sect. 4.3. Overall, the instructors rated the pedagogical effectiveness similarly across all sources. Notably, for the correctness criterion (S2: The given answers are correct), human-crafted questions scored slightly lower than AI-generated ones. This may be due to ambiguous language in human-written foils, particularly in analytical MCQs, where generative AI may outperform humans in producing clearer assertions.

6 Conclusion

We found that a pre-trained LLM can generate valid foils for *analytical multiple-choice questions* regarding a particular topic when a specific learning objective and students' understanding of the topic were given. Our findings suggest that a valid assessment should not only evaluate what the students expected to learn but also identify potential misunderstandings specific to the cohort taking the exam. Our study also demonstrated that an IRT-based analysis can provide a more systematic comparison of AI-generated questions than a survey-based evaluation. These findings highlight critical lessons, emphasizing the need for ethical considerations in evaluation methods as generative AI becomes more widely integrated into authentic learning environments.

This study is limited by the small number of questions analyzed, which constrained the statistical analysis. As IRT requires substantial response data, future work with larger datasets and more instructors is needed to improve generalizability.

Acknowledgments. This research reported here was supported by the National Science Foundation Grant No. 2016966 to North Carolina State University. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

References

1. Gall, M.D.: The use of questions in teaching. Rev. Educ. Res. **40**(5), 707–721 (1970)
2. Ch, D.R., Saha, S.K.: Automatic multiple choice question generation from text: a survey. IEEE Trans. Learn. Technol. **13**(1), 14–25 (2018)
3. Haladyna, T.M., Downing, S.M., Rodriguez, M.C.: A review of multiple-choice item-writing guidelines for classroom assessment. Appl. Measur. Educ. **15**(3), 309–333 (2002)
4. Kurdi, G., et al.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**(1), 121–204 (2020)
5. Leo, J., et al.: Ontology-based generation of medical, multi-term MCQs. Int. J. Artif. Intell. Educ. **29**, 145–188 (2019)
6. Al Faraby, S., Adiwijaya, A., Romadhony, A.: Review on neural question generation for education purposes. Int. J. Artif. Intell. Educ. **34**(3), 1008–1045 (2024)
7. Hwang, K., et al.: Towards automated multiple choice question generation and evaluation: aligning with Bloom's taxonomy. In: International Conference on Artificial Intelligence in Education. Springer (2024)

8. Gorgun, G., Bulut, O.: Instruction-tuned large-language models for quality control in automatic item generation: a feasibility study. Educ. Measur. Issues Pract. (2024)
9. Moore, S., et al.: An automatic question usability evaluation toolkit. In: Olney, A.M., et al. (eds.) Proceedings of the International Conference on Artificial Intelligence in Education, pp. 31–46. Springer (2024)
10. Bulathwela, S., Muse, H., Yilmaz, E.: Scalable educational question generation with pre-trained language models. In: International Conference on Artificial Intelligence in Education. Springer (2023)
11. Elkins, S., et al.: How useful are educational questions generated by large language models? In: International Conference on Artificial Intelligence in Education. Springer (2023)
12. Wang, Z., et al.: Towards human-like educational question generation with large language models. In: International Conference on Artificial Intelligence in Education. Springer (2022)
13. Scaria, N., Dharani Chenna, S., Subramani, D.: Automated educational question generation at different bloom's skill levels using large language models: strategies and evaluation. In: Olney, A.M., et al. (eds.) Proceedings of the International Conference on Artificial Intelligence in Education, pp. 165–179. Springer (2024)
14. Dutulescu, A., et al.: Beyond the obvious multi-choice options: introducing a toolkit for distractor generation enhanced with NLI filtering. In: Olney, A.M., et al. (eds.) Proceedings of the International Conference on Artificial Intelligence in Education, pp. 242–250. Springer (2024)
15. Ch, D.R., Saha, S.K.: Generation of multiple-choice questions from textbook contents of school-level subjects. IEEE Trans. Learn. Technol. **16**(1), 1–13 (2023)
16. Brown, J., Frishkoff, G., Eskenazi, M.: Automatic question generation for vocabulary assessment. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (2005)
17. Correia, R., et al.: Automatic generation of cloze question stems. In: Computational Processing of the Portuguese Language: 10th International Conference, PROPOR 2012, Coimbra, Portugal, 17–20 April 2012. Proceedings 10. Springer (2012)
18. Agarwal, M., Mannem, P.: Automatic gap-fill question generation from text books. In: Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications (2011)
19. Ren, S., Zhu, K.Q.: Knowledge-driven distractor generation for cloze-style multiple choice questions. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)
20. Faizan, A., Lohmann, S.: Automatic generation of multiple choice questions from slide content using linked data. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (2018)
21. Ev, V., Kumar, P.S.: Automated generation of assessment tests from domain ontologies. Semantic Web **8**(6), 1023–1047 (2017)
22. Van Campenhout, R., Hubertz, M., Johnson, B.G.: Evaluating AI-generated questions: a mixed-methods analysis using question data and student perceptions. In: International Conference on Artificial Intelligence in Education. Springer (2022)
23. Liu, M., Rus, V., Liu, L.: Automatic Chinese multiple choice question generation using mixed similarity strategy. IEEE Trans. Learn. Technol. **11**(2), 193–202 (2017)
24. Gherardi, E., et al.: Using knowledge graphs to improve question difficulty estimation from text. In: International Conference on Artificial Intelligence in Education. Springer (2024)
25. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
26. OpenAI, ChatGPT (4o, Dec 2024 version) [Large language model] (2024). https://chat.openai.com/chat
27. Baker, F.B., Kim, S.-H.: Item Response Theory: Parameter Estimation Techniques. CRC Press (2004)

28. Linden, W.J., Hambleton, R.K.: Handbook of Modern Item Response Theory. Volume Two: Statistical Tools. CRC Press (2016)
29. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall, London (1986)

Leveraging Large Reasoning Models for Test Equating Without Anchor Items: A Simulation Study with O1 and DeepSeek-R1

Junlei Du, Qinhua Zheng(✉), and Shuang Li

Research Center of Distance Education, Beijing Normal University, No. 19, Xinjiekouwai Street, Haidian District, Beijing 100875, China
junlei@mail.bnu.edu.cn, {zhengqinhua,lilybnu}@bnu.edu.cn

Abstract. The emergence of advanced reasoning models like O1 and DeepSeek-R1 marks a significant advancement in LLM capabilities, enabling applications in complex tasks such as test equating in educational measurement. This study explored test equating without anchor items by leveraging four large language models—GPT-4o, O1-mini, O1-preview, and DeepSeek-R1—to simulate participant data for linking two distinct mathematics tests (X and Y). The two-parameter item response theory (IRT) model was used for equating. Results showed that advanced reasoning models outperformed general-purpose LLMs, with DeepSeek-R1 demonstrating smaller errors than O1-preview, and both outperforming GPT-4o and O1-mini. Increasing the size of linking groups reduced errors, and prompt characteristics such as gender roles influenced error magnitude. This study highlights the potential of advanced reasoning LLMs in enhancing test equating processes.

Keywords: Test Equating · Large Language Models · Equating Without Anchor Items · Simulation · DeepSeek · O1

1 Introduction

1.1 LLMs as Simulated Participants

Some researchers have conducted simulation experiments and evaluations across multiple disciplines based on this sample generation capability [1–3]. In the fields of sociology and psychometrics, large language models have demonstrated data distributions close to human patterns in areas such as moral judgment [4], economic games [5], and personality testing [6]. For example, Dillon et al. [4] found that GPT-3.5 achieved a 0.95 correlation with human judgments across 464 moral scenarios. Hewitt et al. [5] demonstrated that GPT-4 can effectively mimic human participants in social science surveys and experiments, producing results comparable to real human responses, including attitudes toward policies and vaccination support. Serapio-García et al. [7] noted that upgraded LLMs generate increasingly human-like simulated data. To enhance the realism of LLM-generated outputs, researchers now integrate demographic-based role simulations into

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 135–142, 2025.
https://doi.org/10.1007/978-3-031-99264-3_17

prompts, as shown by Argyle [8], who found that LLM-simulated participants aligned closely with human responses in public opinion surveys. This underscores the importance of examining how prompt-based characteristics influence simulation outcomes.

Thus, it is evident that leveraging LLM-simulated participants for simulation experiments or assessments holds substantial potential. This not only provides a basis for understanding how LLMs comprehend and recognize human psychological characteristics, social behaviors, and decision-making, as well as the differences between these and real human patterns [9], but also promotes the shift towards a human-machine collaborative paradigm in psychological and educational assessment [9].

1.2 Test Equating Without Anchor Items

In educational assessment, equating traditionally uses anchor items to transform scores onto a common scale, as seen in the NEAT Design [10, 11]. However, anchorless equating methods face challenges due to reliance on pre-designed frameworks or impractical additional data collection [12–15]. Marengo proposed a linking group-based equating method to address anchorless equating without pre-design [16]. This approach involves selecting a subset of items from the tests to be equated, forming a new test administered in a low-stakes context, with the resulting sample serving as a linking anchor. While effective, this method incurs significant organizational costs. Segbers introduced a virtual sample approach for vocabulary testing, using simulated performance across age groups as an anchor to enable comparability [17]. Unlike Marengo's method, which relies on real human samples, Segbers' approach uses virtual samples. This distinction suggests that sufficiently advanced AI-generated samples could replace human linking groups. Building on this idea, Du et al. used LLM-simulated samples as linking groups in equating experiments across reading comprehension tests [18]. Their results indicate that GPT-4 outperforms GPT-3.5 in reducing equating errors, with error reduction correlating with increased linking group sample size.

Despite these advancements, limited research has explored the performance of advanced reasoning models, particularly O1 and DeepSeek-R1, in equating contexts. Therefore, this study investigates the anchorless equating performance of four representative LLMs (GPT-4o, O1-mini, O1-preview, and DeepSeek-R1) in mathematics assessments, focusing on their effectiveness in linking group-based anchorless equating. The study addresses the following research questions:

RQ1: Are there significant differences in equating performance among different LLMs?
RQ2: Does the sample size in the linking group significantly affect equating outcomes?
RQ3: Do gender characteristics in prompt design influence equating results?

2 Methods

2.1 Simulation

The study utilized the open-source AQua dataset [19], which contains a diverse range of mathematical problems from simple arithmetic to complex algebra, to construct Test Forms X and Y, each consisting of 100 unique true/false questions. Non-replacement sampling was employed to ensure question uniqueness across forms.

To simulate student responses, four large language models (GPT-4o, O1-mini, O1-preview, and DeepSeek-R1) were tasked with role-playing students. The prompts included randomized student characteristics: gender (male/female), mathematics ability (normally distributed with $\mu = 75$, $\sigma = 10$, capped at 100), and additional demographic details generated using the Faker library. A total of 200 simulated participants responded to either Test Form X or Y, while a connected group answered all 200 questions across both forms. Sample sizes for the connected group were set at 100, 200, and 300 to evaluate equating performance. The prompt template guided responses based on simulated roles, ensuring variability in student profiles while maintaining consistency in task structure.

2.2 IRT Equating Design Based on a Linked Group

This study adopts the linking design equating method proposed by Marengo et al. In traditional designs, due to test-takers' limited cognitive stamina, a fixed proportion of items is typically sampled from different test forms to create new forms. However, simulated participants generated by Large Language Models (LLMs) in the linking design face no such constraints and can respond to all items. The study employs the two-parameter logistic (2PL) model from Item Response Theory (IRT) for equating [20]. For anchor items, the equating relationship between test forms X and Y is given by Eq. 2, where k represents the number of anchor items:

$$P_k(\theta_x, a_x, b_x) = P_k\left(\theta_y, a_y, b_y\right) \tag{1}$$

The connection between forms X and Y via the linking group is established through:

$$\theta_y = \frac{A_{ly}}{A_{lx}}(\theta_x - B_{lx}) + B_{ly} \tag{2}$$

We evaluates equating validity using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which quantify discrepancies between equated parameters and true parameter estimates [21].

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\left|\widehat{\varsigma}_r - \varsigma_r\right| \tag{3}$$

$$RMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\varsigma}_r - \varsigma_r\right)^2} \tag{4}$$

Here, $N = 20$ represents 20 replications for each condition (varying LLMs, gender, and linking group sample sizes), R is the number of items, $\widehat{\varsigma}_r$ denotes equated parameters, and ς_r represents true parameter values. MAE and RMSE are computed separately for discrimination a and difficulty b parameters to assess equating effectiveness. All simulation procedures in this study were implemented in Python using specialized libraries. In existing IRT-based equating research, MAE and RMSE typically range between 0.2 and 0.8, with lower values indicating superior equating performance [22, 23]. The integration of Large Language Models (LLMs) was conducted using OpenAI's API or DeepSeek's API, while Item Response Theory (IRT) operations, including parameter estimation and equating, were executed via the girth library.

3 Results

3.1 Descriptive Statistical Analysis of Simulated Data

Descriptive statistical analysis of simulated participant performance on Test Form X and Test Form Y was conducted using metrics such as mean, standard deviation, skewness, and kurtosis for different large language models.

Table 1. Descriptive Statistical Analysis Results

LLMs	Test Form X				Test Form Y			
	Mean	Std	Sk	Ku	Mean	Std	Sk	Ku
GPT-4o	43.03	4.99	0.11	−0.15	44.80	4.80	0.02	0.14
O1-mini	39.96	4.57	0.29	0.49	43.40	5.44	−0.47	−0.02
O1-preview	49.11	4.80	−0.65	2.04	50.31	5.67	−1.27	3.71
DeepSeek-R1	63.47	15.24	0.06	−1.71	62.67	14.23	−0.09	−1.42

Table 1 provides descriptive statistics for simulated test responses across four large language models (LLMs). The results indicate that OpenAI models (GPT-4o, O1-mini, and O1-preview) systematically underestimated the simulated ability level, with mean scores ranging from 43 to 50 and standard deviations clustered around 5, reflecting low dispersion and limited diversity. In contrast, DeepSeek-R1 demonstrated superior performance with a mean score of approximately 63 and standard deviations exceeding 10, suggesting greater variability and flatter score distributions compared to the OpenAI series.

3.2 Comparison of Different LLMs

Table 2 compares equating errors across LLMs using MAE and RMSE for discrimination (a) and difficulty (b) parameters. The analysis reveals distinct hierarchical patterns in equating parameter error across different LLMs, with DeepSeek-R1 emerging as the model with the lowest error.

DeepSeek-R1 demonstrated superior equating performance across both discrimination (a) and difficulty (b) parameters, outperforming GPT-4o by 7–8% in error reduction while maintaining a clear performance gradient over O1-mini and O1-preview. This trend highlights that the advancements in LLM architecture, training methodologies, and logical reasoning capabilities collectively contribute to reduced measurement equating error.

Table 2. Error Results of Test Equating Parameters for Different LLMs

LLMs	*a*		*b*	
	MAE	RMSE	MAE	RMSE
GPT-4o	0.650	0.798	0.731	0.930
O1-mini	0.634	0.774	0.706	0.903
O1-preview	**0.615**	**0.747**	**0.688**	**0.876**
DeepSeek-R1	**0.598**	**0.731**	**0.673**	**0.858**

3.3 Comparison of Different Link Group Sizes

In this section, we compared the error results of discrimination (a) and difficulty (b) parameters across different link group sizes (100, 200, and 300) (Table 3).

Table 3. Error Results of (*a*) and (*b*) for Different Link Group Sizes

IRT parameters	LLMs	MAE			RMSE		
		Link Group Size			Link Group Size		
		100	200	300	100	200	300
a	GPT-4o	0.658	0.657	**0.642**	0.806	0.804	**0.791**
	O1-mini	0.641	0.636	**0.617**	0.786	0.778	**0.738**
	O1-preview	0.637	0.624	**0.580**	0.83	0.740	**0.697**
	DeepSeek-R1	0.621	0.602	**0.583**	0.766	0.749	**0.716**
b	GPT-4o	0.741	0.737	**0.723**	0.941	0.938	**0.925**
	O1-mini	0.714	0.707	**0.699**	0.913	0.912	**0.889**
	O1-preview	0.704	0.685	**0.667**	0.897	0.873	**0.852**
	DeepSeek-R1	0.678	0.673	**0.662**	0.872	0.862	**0.841**

The analysis of discrimination (a) and difficulty (b) parameter errors across varying sample sizes revealed distinct sensitivities among models. O1-preview demonstrated the most pronounced improvements, with significant reductions in MAE and RMSE as sample sizes increased, indicating heightened sensitivity to sample size. In contrast, DeepSeek-R1 exhibited stable optimization, maintaining robust parameter estimation even with smaller samples, while GPT-4o showed minimal improvements beyond N = 200. O1-mini exhibited moderate declines in error metrics, suggesting potential for further error correction with larger sample sizes. Overall, O1-preview outperformed other models in sensitivity to sample size, while DeepSeek-R1 provided consistent and reliable parameter estimation across sample sizes.

3.4 Comparison of Different Genders

In this section, we compared the error results of discrimination (a) and difficulty (b) parameters across different genders (Table 4).

Table 4. Error Results of Test Equating Parameters for Genders

LLMs	Gender	*a*		*b*	
		MAE	*RMSE*	*MAE*	*RMSE*
GPT-4o	Male	0.646	0.789	0.730	0.925
	Female	0.647	0.801	0.723	0.921
O1-mini	Male	0.650	0.797	0.714	0.908
	Female	0.628	0.773	0.698	0.894
O1-preview	Male	0.634	0.753	0.683	0.865
	Female	0.601	0.712	0.698	0.890
DeepSeek-R1	Male	0.573	0.679	0.669	0.848
	Female	0.612	0.745	0.682	0.869

The error results of test equating parameters across gender characteristics revealed distinct patterns among different models. GPT-4o demonstrated gender-neutral performance, with minimal discrepancies in parameter errors between male and female groups (within 1.6%). O1-mini and O1-preview showed pronounced female advantages, particularly in parameter a, while DeepSeek-R1 exhibited a notable male advantage in parameter a and minimal gender differences in parameter b. These findings highlight model-specific gender-related variations in parameter estimation accuracy.

4 Conclusion

This study evaluated four LLMs (GPT-4o, O1-mini, O1-preview, DeepSeek-R1) in anchorless test equating. Results showed that DeepSeek-R1 exhibited superior accuracy in parameter estimation (7.9% lower MAE) due to its advanced reasoning and probabilistic handling. Meanwhile, the O1 models demonstrated high sensitivity to sample size, with a 9.3% MAE reduction at $N = 300$. The findings indicated that model selection should prioritize simulation fidelity over raw ability. Generally, larger link groups reduced equating errors, but diminishing returns necessitated cost-accuracy trade-offs. Gender-based disparities in prompts were also observed, such as DeepSeek-R1's male bias and O1's female advantage, which highlighted latent training-data biases and suggested that role-playing characteristics should be comprehensively considered in prompt engineering. Limitations of the study included oversimplified role simulations, modest sample sizes for high-dimensional models, and a narrow focus on IRT-based equating. Future research should explore multimodal feature simulations and cross-methodological validation to optimize LLM applications in psychometrics.

References

1. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–22. Association for Computing Machinery, New York (2023)
2. Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: Proceedings of the 40th International Conference on Machine Learning, pp. 337–371. PMLR (2023)
3. Gao, C., et al.: Large language models empowered agent-based modeling and simulation: a survey and perspectives. Humanit. Soc. Sci. Commun. **11**, 1–24 (2024). https://doi.org/10.1057/s41599-024-03611-3
4. Dillion, D., Tandon, N., Gu, Y., Gray, K.: Can AI language models replace human participants? Trends Cogn. Sci. **27**, 597–600 (2023). https://doi.org/10.1016/j.tics.2023.04.008
5. Hewitt, L., Ashokkumar, A., Ghezae, I., Willer, R.: Predicting Results of Social Science Experiments Using Large Language Models (2024)
6. Petrov, N.B., Serapio-García, G., Rentfrow, J.: Limited Ability of LLMs to Simulate Human Psychological Behaviours: a Psychometric Analysis (2024)
7. Serapio-García, G., et al.: Personality Traits in Large Language Models (2023)
8. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: using language models to simulate human samples. Polit. Anal. **31**, 337–351 (2023). https://doi.org/10.1017/pan.2023.2
9. Leng, Y., Yuan, Y.: Do LLM Agents Exhibit Social Behavior? (2024)
10. Skaggs, G., Lissitz, R.W.: IRT test equating: relevant issues and a review of recent research. Rev. Educ. Res. **56**, 495–529 (1986). https://doi.org/10.3102/00346543056004495
11. Kim, S., von Davier, A.A., Haberman, S.: An alternative to equating with small samples in the non-equivalent groups anchor test design. ETS Res. Rep. Ser. **2006**, i–40 (2006). https://doi.org/10.1002/j.2333-8504.2006.tb02033.x
12. Liao, C.-W., Livingston, S.A.: Examining an alternative to score equating: a randomly equivalent forms approach. ETS Res. Rep. Ser. **2008**, i–22 (2008). https://doi.org/10.1002/j.2333-8504.2008.tb02100.x
13. Livingston, S.A.: Demographically adjusted groups for equating test scores. ETS Res. Rep. Ser. **2014**, 1–10 (2014). https://doi.org/10.1002/ets2.12030
14. Liu, Q., et al.: Fuzzy cognitive diagnosis for modelling examinee performance. ACM Trans. Intell. Syst. Technol. **9**, 48:1–48:26 (2018). https://doi.org/10.1145/3168361
15. Tatsuoka, K.K.: Rule space: an approach for dealing with misconceptions based on item response theory. J. Educ. Meas. **20**, 345–354 (1983)
16. Marengo, D., Miceli, R., Rosato, R., Settanni, M.: Placing multiple tests on a common scale using a post-test anchor design: effects of item position and order on the stability of parameter estimates. Front. Appl. Math. Stat. **4** (2018). https://doi.org/10.3389/fams.2018.00050
17. Segbers, J., Schroeder, S.: How many words do children know? A corpus-based estimation of children's total vocabulary size. Lang. Test. **34**, 297–320 (2017). https://doi.org/10.1177/0265532216641152
18. Du, J., Zheng, Q., Song, Y.: The Application of Large Language Models for Test Equating with No Common Items: A Case Study on Reading Literacy Assessment. zggx 1–8 (2024). https://doi.org/10.14188/j.1671-8836.2024.0088
19. Ling, W., Yogatama, D., Dyer, C., Blunsom, P.: Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems (2017)
20. Haebara, T.: Equating logistic ability scales by a weighted least squares method. Jpn. Psychol. Res. **22**, 144–149 (1980). https://doi.org/10.4992/psycholres1954.22.144

21. Yao, L.: Multidimensional linking for domain scores and overall scores for nonequivalent groups. Appl. Psychol. Meas. **35**, 48–66 (2011). https://doi.org/10.1177/0146621610373095
22. He, Y., Cui, Z., Osterlind, S.J.: New robust scale transformation methods in the presence of outlying common items. Appl. Psychol. Meas. **39**, 613–626 (2015). https://doi.org/10.1177/0146621615587003
23. Wang, S., Zhang, M., You, S.: A comparison of IRT observed score kernel equating and several equating methods. Front. Psychol. **11** (2020). https://doi.org/10.3389/fpsyg.2020.00308

Leveraging Knowledge Profiles and Generative AI for Realistic Student Response Generation

Eylül Ipçi, Tanya Nazaretsky(✉), and Tanja Käser

EPFL, Lausanne, Switzerland
{eylul.ipci,tanya.nazaretsky,tanja.kaeser}@epfl.ch

Abstract. The scarcity of high-quality labeled data often hinders the effective use of automated formative feedback in education. While analytic rubrics offer a reliable framework for automated grading, training robust models still requires hundreds of expert-labeled responses, an expensive and time-consuming process. This paper proposes a methodology for generating diverse, rubric-aligned synthetic student responses using large language models (LLMs) guided by knowledge profiles and representative examples. We introduce two profile-based generation strategies, straightforward and error-informed, and evaluate them compared to a dataset (N = 585) of authentic open-ended logical-proof responses from a Discrete Mathematics course. We analyze the diversity and realism of the generated datasets using embedding-based distance metrics and PCA and assess their utility for training automated grading models. Our results show that synthetic responses are less diverse than authentic ones, and models trained solely on generated data perform worse than those trained on real data. However, combining small authentic datasets with generated data significantly improves model performance, suggesting as an effective augmentation strategy in low-resource educational settings.

Keywords: Generative AI · Data Augmentation · Higher Education · Formative Feedback

1 Introduction

Feedback is one of the most important components of the effective learning process [25]. However, in higher education, personalized feedback delivery is often absent or delayed due to the large and diverse student population [6]. Recently, advances in Natural Language Processing (NLP) utilizing novel deep learning approaches and state-of-the-art language models (e.g., BERT, RoBERTa, GPT, Llama, etc.) are a promising direction in automating feedback provision [15,29]. Previous research has employed zero- and few-shot generative AI techniques to generate formative feedback [5,10,22,26]. However, this approach often suffers from issues such as lack of transparency, inconsistency, and lower quality compared to human feedback.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 143–151, 2025.
https://doi.org/10.1007/978-3-031-99264-3_18

One effective way to structure formative feedback creation is using analytic rubrics [12,14,19,23,28]. Indeed, task-dependent analytic grading rubrics offer a structured and straightforward approach to formative assessment. Such analytic rubrics consist of several categories, each corresponding to a specific learning objective, such as a particular aspect, idea, concept, or step that should be included in a correct response or an error or misconception that should be avoided [2]. Categories are typically binary, where a score of 1 indicates that the corresponding aspect is adequately addressed, and a score of 0 indicates it is missing or incorrectly addressed [1]. Such detailed and structured evaluation provides a robust framework for formative assessment [4,18,21,24,27].

Systems that employ analytic rubrics as part of the automated feedback generation process have been reported to be more reliable, transparent, and less biased [3,13,21,30]. However, these systems typically utilize supervised Machine Learning (ML) techniques and, as such, require relatively large datasets ($\sim$ 500-600 responses to achieve Kappa of 0.6 [3,28,31]), which are accurately labeled by human experts, which is challenging to collect in educational settings. To address the challenge of labeled data scarcity, NLP research has introduced innovative strategies, including text augmentation, knowledge distillation, and zero- and few-shot learning [7,17]. These augmentation techniques have been successfully applied in educational settings [32]. For instance, [8] employed four methods: word-level masked language models, noise injection, hypo/hypernym substitution, and oversampling. [20] used two methods: heuristics-based rules and back-translation-based paraphrasing, to generalize models from Hebrew to Turkish. Additionally, [9] utilized GPT-3.5 to augment minority classes in grading responses in the context of causal reasoning with small, unbalanced datasets in English and French. [11] utilized GPT-4 to augment datasets with minority class examples for assessment items requiring students to explain scientific phenomena. [16] utilized GPT-3.5 to generate synthetic student responses with specific pre- and misconceptions in physics quantitative reasoning tasks, thereby augmenting the minority classes. However, none of the above studies evaluated datasets that consisted of generated responses only.

This study addresses the scarcity issue in labeled educational datasets by presenting a methodology for generating diverse and realistic student responses based on task, analytic rubric, and several examples of authentic student answers only. To this end, we answer the following two research questions: [RQ1] To what extent are the generated responses similar to the original student responses? [RQ2] Can the generated responses be used to train high-quality small language models for automated grading of original student responses?

2 Methodology

Our response generation and evaluation pipeline is presented in Fig. 1 with each stage presented below in more detail.

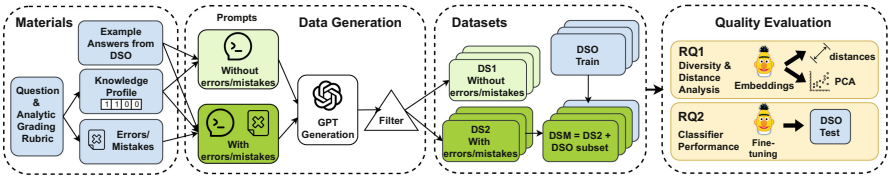


Fig. 1. The research pipeline

2.1 Materials

The DSO dataset consisted of 585 student responses to an open-ended logical proof question, formulated as follows: "Prove by contrapositive that if x^2 is irrational, then x must also be irrational." The question was part of a graded quiz from the Discrete Mathematics course offered to first-year undergraduate students in the Computer Science faculty at EPFL. The ethics committee of EPFL approved the study (HREC 017-2023). Student responses were graded by the pedagogical team as part of the course's routine assessment using four binary rubrics: [R1] formulation of the contraposition rule, [R2] applying it to a concrete case, [R3] detailed logical proof, and [R4] conclusion. Each rubric was accomplished by a correct solution, a list of errors (which caused grade reduction), and a list of small mistakes (which did not influence the grade). More details are presented in Supplementary Materials.

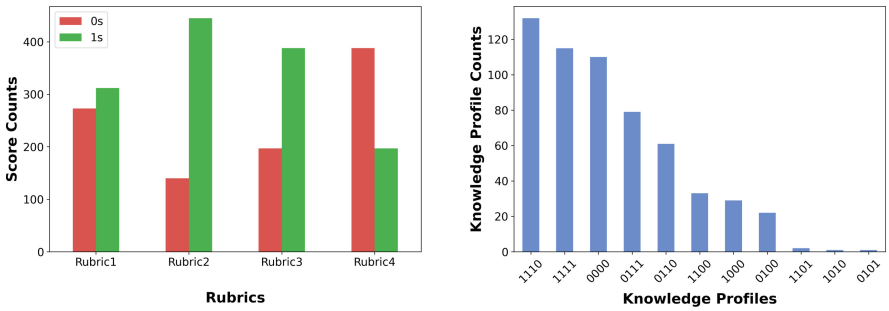


Fig. 2. The frequencies of the scores (0/1) per rubric and the frequencies of the knowledge profiles

The **knowledge profile** is defined as the combination of binary scores assigned to the four rubrics used in grading, resulting in a unique four-digit binary vector that represents the student's understanding levels:

– **Profiles 0000, 1000**: Lack of understanding of proof by contraposition and ability to apply it to specific task cases.

- **Profiles 1100, 0100**: Basic Knowledge of how to apply the contraposition rule to a specific case, but the proof is missing or incorrect.
- **Profiles 0110, 1110, and 0111**: Substantial knowledge of contraposition and its application, but proof is incomplete or unstructured.
- **Profile 1111**: Substantial knowledge of contraposition and its application.

The frequencies of the knowledge profiles are presented in Fig. 2.

2.2 Data Generation

Prompt Design. In this project, we utilized the OpenAI GPT-4o interface. The prompt structure reflected the role-task-format (RTF) and was designed to include the following components: [System Role] You are a text generator; [Task Definition] Generate a first-year CS student's answer to the contraposition question, making a particular mistake; [Output Format] Use the following examples. The code and the generated examples are in Supplementary Materials.

Datasets Generation. In this study, to create a balanced and realistic dataset, we utilized five knowledge profiles (0000, 1000, 0110, 1110, and 0111) representing a range of possible student knowledge levels. Using these profiles, we generated two synthetic student response datasets, each of size N = 500. One dataset (DS1) was created with a prompt that did not specify a particular mistake or error. In contrast, the other (DS2) was created by providing the error or mistake that should be present in the generated synthetic response. In this case, the errors were carefully paired with corresponding knowledge profiles. For example, in the case of profile 0111, only errors related to Rubric 1 were allowed, reflecting its zero score in that rubric. For DS2, each possible combination of profile with error/mistake was used to generate 20–30 synthetic responses. For DS1, we generated 150 responses per profile. Next, we manually filtered out responses that deviated from the required criteria, specifically those unrelated to a specific profile or error. Finally, for each dataset, we randomly chose 100 responses per each profile, resulting in two datasets of size 500.

2.3 Analytic Approach

RQ1: Diversity and Distance Analysis. In the context of RQ1, to evaluate the quality of the data we generated (DS1, DS2), we used two measures: *diversity within a dataset* and *distance between two datasets*. Our goal is to generate data with a diversity value similar to the original dataset (DSO) and the distance value between the generated and original datasets to be smaller. The diversity of dataset $X = \{x_1, x_2, ...x_n\}$ was quantified as $DIV(X) = \frac{1}{n^2} \sum_{j,k} distance(x_j, x_k)$. The distance between two datasets $X = \{x_1, x_2, ...x_n\}$ and $Y = \{y_1, y_2, ...x_m\}$ was quantified as $DIS(X, Y) = \frac{1}{nm} \sum_{j,k} distance(x_j, y_k)$. In both cases, we used Euclidean and Cosine distances, where Cosine distance was defined as 1 - Cosine Similarity. The textual response

data was converted to DistilBERT embeddings by extracting the [CLS] token. The diversity was calculated per dataset (DSO, DS1, and DS2) for the entire dataset and for each knowledge profile. The distance was calculated for DS1 and DS2 compared to DSO for the entire dataset and per knowledge profile. We used PCA to visualize the embedding distributions and gain an intuitive understanding of the diversity of the generated data relative to the original responses for the entire dataset and per knowledge profile.

RQ2: Classifier Performance. To answer RQ2, we trained the DistilBERT classifiers per rubrics (R1, R2, R3, R4) and datasets (DSO, DS1, DS2, the increasing subsets of DSO, and several combinations of DS2 and increasing subsets of DSO). Next, we tested them on the held-out part (25%) of DSO (DSO-Test). Hyperparameters were optimized via 5-fold cross-validation. Once the optimal settings were determined, each classifier was trained for 14 epochs with a batch size of 16 using the AdamW optimizer (learning rate 2×10^{-5}, weight decay 0.01). The classifiers trained on DSO served as the baseline models.

3 Results

RQ1: Diversity and Distance. The results of our diversity and distance analysis are presented in Table 1 and Fig. 3. The diversity analysis confirmed that generated responses were less diverse than original authentic student responses for both generations methods (DS1 - prompting per profile without specifying errors and mistakes, and DS2 - by providing information about specific mistakes in the prompt). Deeper analysis, including PCA visualizations, revealed a certain overlap between the clouds of points corresponding to the original and generated responses for both DS1 and DS2, with a significant portion of the original

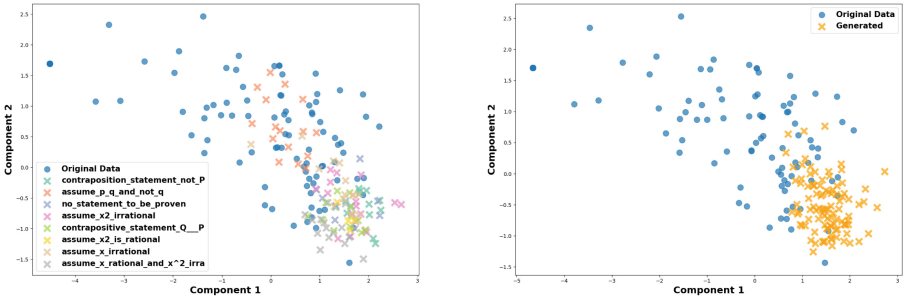


Fig. 3. PCA analysis for 0000 knowledge profile. The blue points represent the original dataset (DSO). The ×-s represent generated points from DS1 (right-hand side) and DS2 (left-hand side). For DS2, the colors of the ×-s represent different errors and mistakes used in the prompt for generating the corresponding responses (Color figure online)

Table 1. Diversity and distance measures across profiles (0000, 1000, 0110, 1110, 0111) and the entire dataset, using Cosine and Euclidean distances

	0000		1000		0110		1110		0111		Entire Dataset	
Datasets	Cos	Eucl.	Cos	Eucl.	Cos	Eucl.	Cos	Eucl.	Cos	Eucl.	Cos	Eucl.
DIV(DSO)	0.099	5.299	0.041	3.660	0.044	3.597	0.032	3.195	0.037	3.396	0.066	4.144
DIV(DS1)	0.024	2.827	0.022	2.766	0.018	2.450	0.013	2.068	0.015	2.101	0.024	2.796
DIV(DS2)	0.026	2.935	0.023	2.788	0.018	2.492	0.015	2.230	0.014	2.157	0.023	2.772
DIS(DSO, DS1)	0.088	5.055	0.041	3.678	0.044	3.711	0.027	2.981	0.036	3.401	0.050	3.889
DIS(DSO, DS2)	0.084	4.907	0.044	3.802	0.042	3.627	0.029	3.074	0.035	3.376	0.050	3.884

Table 2. DistilBERT Classifiers Performance. DSO[N] denotes DSO subset of size N

	Rubric 1			Rubric 2			Rubric 3			Rubric 4		
Datasets	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa	Acc	F1	Kappa
DSO	0.864	0.864	0.726	0.898	0.897	0.665	0.871	0.869	0.694	0.844	0.847	0.664
DS1	0.796	0.790	0.583	0.640	0.678	0.276	0.708	0.713	0.360	0.674	0.554	0.014
DS2	0.850	0.850	0.700	0.728	0.756	0.386	0.708	0.717	0.411	0.674	0.565	0.028
DSO[100]	0.789	0.786	0.572	0.905	0.900	0.664	0.878	0.873	0.702	0.782	0.783	0.510
DSO[200]	0.823	0.823	0.646	0.905	0.905	0.691	0.891	0.889	0.741	0.816	0.818	0.593
DS2 + DSO[100]	0.837	0.837	0.672	0.898	0.901	0.690	0.857	0.856	0.666	0.789	0.770	0.460
DS2 + DSO[200]	0.864	0.864	0.727	0.891	0.895	0.674	0.898	0.899	0.772	0.823	0.816	0.570

points not covered by the generated points. An example of such an analysis for knowledge profile 0000 is presented in Fig. 3. Comparing the left-hand (DS2) and right-hand (DS1) side figures, we conclude that adding information about specific mistakes increased the diversity, and the corresponding cloud of points (DS2) overlaps more with the original responses cloud than DS1. However, still a significant area of original points is covered by none of the generation methods.

RQ2: Classifier Performance. The performance results of all classifiers are presented in Table 2. The classifiers trained on the original dataset (DSO) achieved sufficient performance (Table 2). Decreasing the DSO size influences the classifier's performance for rubrics R1 and R4 (Kappa < 0.6), while for rubrics R2 and R3, even 100 responses were sufficient (Kappa > 0.6). While the classifiers trained on the dataset generated using specific error/mistake information (DS2) achieved better performance than those trained on DS1 (with knowledge profile information only), they still did not achieve a satisfactory performance for all profiles besides R1. However, combining the generated data with a certain number of original responses tended to improve the performance of the corresponding classifiers besides Rubric R4, where the results were inconsistent.

4 Conclusions

Consistent with previous studies [3,28,31], our results confirm that approximately 500 responses are sufficient for training reliable DistilBERT classifiers for all rubrics (Kappa > 0.6), despite the dataset's unbalanced nature for R2, R3, and R4. However, the classifiers trained on the generated datasets of the same size failed to achieve similar performance quality, except for DS2 with Rubric R1. We attribute this finding to the lower diversity of the generated responses compared to authentic student data, limiting the classifiers' ability to capture the complete variety of correct and incorrect formulations. This observation aligns with prior research emphasizing the importance of diversity in training data for robust model generalization.

Our results suggest that in low-resource settings–with only 100–200 labeled responses available–targeted data generation can serve as a valuable augmentation strategy, significantly improving classifier performance when combined with original data. In future work, we aim to cover more knowledge profiles and produce larger datasets, apply our method using embeddings from fine-tuned LLM models, test our approach on other types of reasoning tasks, and test our approach effectiveness in real learning settings with EPFL students.

Acknowledgements. This study was funded by the Swiss State Secretariat for Education, Research and Innovation SERI.

References

1. Aleven, V., Mclaren, B., Roll, I., Koedinger, K.: Toward meta-cognitive tutoring: a model of help seeking with a cognitive tutor. Int. J. Artif. Intell. Educ. **16**(2), 101–128 (2006)
2. Allen, D., Tanner, K.: Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. CBE-Life Sci. Educ. **5**(3), 197–203 (2006)
3. Ariely, M., Nazaretsky, T., Alexandron, G.: Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. Int. J. Artif. Intell. Educ. **33**(1), 1–34 (2023)
4. Ariely, M., Nazaretsky, T., Alexandron, G.: Causal-mechanical explanations in biology: applying automated assessment for personalized learning in the science classroom. J. Res. Sci. Teach. (2024)
5. Bewersdorff, A., Zhai, X., Roberts, J., Nerdel, C.: Myths, mis- and preconceptions of artificial intelligence: a review of the literature. Comput. Educ. Artif. Intell. **4**, 100143 (2023)
6. Boud, D., Molloy, E.: Rethinking models of feedback for learning: the challenge of design. Assess. Eval. High. Educ. **38**, 698–712 (2013)
7. Chen, H., Saha, A., Hoi, S., Joty, S.: Personalised distillation: Empowering open-sourced LLMs with adaptive learning for code generation. arXiv preprint arXiv:2310.18628 (2023)

8. Cochran, K., Cohn, C., Hutchins, N., Biswas, G., Hastings, P.: Improving automated evaluation of formative assessments with text data augmentation. In: International Conference on Artificial Intelligence in Education, pp. 390–401 (2022)
9. Cochran, K., Cohn, C., Rouet, J.F., Hastings, P.: Improving automated evaluation of student text responses using GPT-3.5 for text data augmentation. In: International Conference on Artificial Intelligence in Education, pp. 217–228 (2023)
10. Dai, W., et al.: Can large language models provide feedback to students? A case study on ChatGPT. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 323–325. IEEE (2023)
11. Fang, L., Lee, G.G., Zhai, X.: Using GPT-4 to augment unbalanced data for automatic scoring. arXiv preprint arXiv:2310.18365 (2023)
12. Jescovitch, L.N., et al.: Deconstruction of holistic rubrics into analytic rubrics for large-scale assessments of students' reasoning of complex science concepts. Pract. Assess. Res. Eval. **24**(1), 7 (2019)
13. Jescovitch, L.N., et al.: Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. J. Sci. Educ. Technol. **30**(2), 150–167 (2021)
14. Jonsson, A., Svingby, G.: The use of scoring rubrics: reliability, validity and educational consequences. Educ. Res. Rev. **2**(2), 130–144 (2007)
15. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
16. Kieser, F., Wulff, P., Kuhn, J., Küchemann, S.: Educational data augmentation in physics education research using ChatGPT. Phys. Rev. Phys. Educ. Res. **19**(2), 020150 (2023)
17. Li, B., Hou, Y., Che, W.: Data augmentation approaches in natural language processing: a survey. Ai Open **3**, 71–90 (2022)
18. Liu, O.L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., Linn, M.C.: Automated scoring of constructed-response science items: prospects and obstacles. Educ. Meas. Issues Pract. **33**(2), 19–28 (2014)
19. Moskal, B.M., Leydens, J.A.: Scoring rubric development: validity and reliability. Pract. Assess. Res. Eval. **7**(1), 10 (2019)
20. Nazaretsky, T., Yolcu, H., Ariely, M., Alexandron, G.: Towards automated assessment of scientific explanations in Turkish using language transfer. In: Proceedings of the 16th International Conference on Educational Data Mining, pp. 453–457 (2023)
21. Nehm, R.H., Ha, M., Mayfield, E.: Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J. Sci. Educ. Technol. **21**, 183–196 (2012)
22. Pankiewicz, M., Baker, R.S.: Large language models (GPT) for automating feedback on programming assignments. arXiv preprint arXiv:2307.00150 (2023)
23. Popham, W.J.: What's wrong-and what's right-with rubrics. Educ. Leadersh. **55**, 72–75 (1997)
24. Rahimi, Z., Litman, D., Correnti, R., Wang, E., Matsumura, L.C.: Assessing students' use of evidence and organization in response-to-text writing: using natural language processing for rubric-based automated scoring. Int. J. Artif. Intell. Educ. **27**(4), 694–728 (2017)
25. Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. **78**, 153–189 (2008)
26. Steiss, J., et al.: Comparing the quality of human and ChatGPT feedback of students' writing. Learn. Instr. **91**, 101894 (2024)

27. Uhl, J.D., Sripathi, K.N., Meir, E., Merrill, J., Urban-Lurain, M., Haudek, K.C.: Automated writing assessments measure undergraduate learning after completion of a computer-based cellular respiration tutorial. CBE–Life Sci. Educ. **20**(3), ar33 (2021)
28. Wilson, C.D., et al.: Using automated analysis to assess middle school students' competence with scientific argumentation. J. Res. Sci. Teach. **61**(1), 38–69 (2024)
29. Yan, L., et al.: Practical and ethical challenges of large language models in education: a systematic scoping review. Br. J. Edu. Technol. **55**(1), 90–112 (2024)
30. Yune, S.J., Lee, S.Y., Im, S.J., Kam, B.S., Baek, S.Y.: Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students. BMC Med. Educ. **18**, 1–6 (2018)
31. Zhai, X., Shi, L., Nehm, R.H.: A meta-analysis of machine learning-based science assessments: factors impacting machine-human score agreements. J. Sci. Educ. Technol. **30**, 361–379 (2021)
32. Neshaei, S.P., Davis, R.L., Mejia-Domenzain, P., Nazaretsky, T., Käser, T.: Bridging the data gap: using LLMs to augment datasets for text classification. In: Proceedings of the 18th International Conference on Educational Data Mining (2025)

LLMs to Support K12 Teachers in Culturally Relevant Pedagogy: An AI Literacy Example

Jiayi Wang[1(✉)], Ruiwei Xiao[2(✉)], Xinying Hou[3], Hanqi Li[4], Ying Jui Tseng[2], John Stamper[2], and Kenneth Koedinger[2]

[1] Northwestern University, Evanston, IL 60208, USA
jiayiwang2025@u.northwestern.edu
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
{ruiweix,yingjuit,jstamper,koedinger}@cmu.edu
[3] University of Michigan, Ann Arbor, MI 48109, USA
xyhou@umich.edu
[4] New York University, New York, NY 10003, USA
hl4893@nyu.edu

Abstract. Culturally Relevant Pedagogy (CRP) is vital in K-12 education, yet teachers struggle to implement CRP into practice due to time, training, and resource gaps. This study explores how Large Language Models (LLMs) can address these barriers by introducing CulturAIEd, an LLM tool that assists teachers in adapting AI literacy curricula to students cultural contexts. Through an exploratory pilot with four K-12 teachers, we examined CulturAIEds impact on CRP integration. Results showed CulturAIEd enhanced teachers confidence in identifying opportunities for cultural responsiveness in learning activities and making culturally responsive modifications to existing activities. They valued CulturAIEd's streamlined integration of student demographic information, immediate actionable feedback, which could result in high implementation efficiency. This exploration of teacher-AI collaboration highlights how LLM can help teachers include CRP components into their instructional practices efficiently, especially in global priorities for future-ready education, such as AI literacy.

Keywords: Culturally Relevant Pedagogy · AI Literacy · Large Language Models · K-12 Education · Teacher Support

1 Introduction

Culturally Relevant Pedagogy (CRP), rooted in Ladson-Billings' foundational framework, emphasizes centering students cultural identities as assets in education [17]. CRP has been successfully implemented in traditional subjects such as mathematics, history, and language arts [5,14]. Despite benefits like improved engagement, achievement, and self-esteem [8,10], teachers struggle with CRP due

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 152–160, 2025.
https://doi.org/10.1007/978-3-031-99264-3_19

to limited time, training, and discomfort with cultural topics [24]. Concurrently, the rise of artificial intelligence (AI) in society has spurred calls for K-12 AI literacy initiatives [25]. However, there is a notable lack of culturally responsive AI literacy learning resources that address the needs of a diverse population of learners. This gap exists because AI literacy is an emerging subject with limited educational resources tailored to different cultural and demographic contexts [11,19,20]. This dual challenge, integrating CRP into AI education, presents a critical opportunity for innovation. Large Language Models (LLMs) have shown promise in streamlining lesson planning and content adaptation [23], but their potential to support CRP remains underexplored.

To address this gap, we present CulturAIEd, an LLM-powered system designed to help teachers contextualize AI literacy activities through CRP principles. CulturAIEd combines demographic characteristics of the students, rubric-based feedback, and just-in-time coaching to simplify the adaptation process while encouraging meaningful cultural integration. This paper investigates two research questions: **RQ1:** How do LLM-supported tools like CulturAIEd influence teachers ability to design culturally responsive AI literacy activities? **RQ2:** What are the perceived strengths and opportunities of such tools in addressing CRP implementation barriers? Through a pilot with four K12 teachers, conducted in preparation for future mixed-methods study, we explored how CulturAIEd influences their confidence, efficiency, and pedagogical strategies. Our findings highlight the tools role in reducing planning burdens and fostering deeper engagement with CRP. To our knowledge, this work represents the first effort to empower K-12 educators teaching traditional subjects to design culturally responsive AI literacy activities with low effort using LLMs. By situating AI as a partner in CRP, this work advances discussions on equitable technology integration and teacher empowerment in diverse classrooms.

2 Related Work

Culturally Relevant Pedagogy (CRP) emphasizes leveraging students' cultural backgrounds as key assets in education. Ladson-Billings' foundational work defines CRP through promoting academic success, developing cultural competence, and fostering sociopolitical consciousness [17]. Empirical evidence consistently highlights CRP's positive impact on academic achievement, student engagement, attendance, motivation, social-emotional development and ethnic-racial identity [8,10,15]. Nonetheless, teachers often encounter significant challenges implementing CRP, including constraints like limited time, curricular rigidity, lack of culturally relevant resources, and discomfort addressing sensitive cultural topics [21,24]. Building upon CRP, Geneva Gay introduced the closely aligned concept of Culturally Responsive Teaching (CRT), emphasizing more explicit strategies for operationalizing cultural responsiveness within daily teaching practices [13]. CRT particularly informs practical methodologies related to modifying course materials, which is perceived as challenging by many teachers. Teachers frequently require administrative support, professional training, and accessible tools to effectively integrate CRP/CRT methods.

As AI becomes integral to societal infrastructure, there is increasing recognition of the necessity for K-12 AI literacy education [9]. However, integrating AI into curricula remains exploratory and lacks systematic frameworks [9]. Current AI literacy programs largely employ a generalized approach, minimally tailored to specific cultural contexts, despite evidence supporting the effectiveness of contextualized content [12]. Scholars advocate for modular, culturally adaptable AI curricula to better meet the diverse backgrounds and needs of students [9].

LLMs have emerged as promising tools to alleviate challenges in implementing CRP/CRT practices. LLMs support teachers by significantly reducing time spent on lesson planning and content adaptation [16]. Early studies indicate that teachers can utilize LLMs to create lesson materials and adapt existing curricula, addressing key barriers like limited resources and heavy workloads [6,23]. However, most AI-driven teacher tools to date focus on general lesson planning or content generation rather than explicitly supporting cultural relevance. Some tools address content management, sustainability goals, or improving lesson plans, reflecting a growing interest in AI for education [22], but without directly incorporating CRP frameworks. CulturAIEd is novel in that it integrates a CRT checklist [4] and demographic customization into the LLMs generative process, providing feedback layered with cultural pedagogical guidance. We describe CulturAIEds design and a pilot study evaluating its use below.

3 Methods

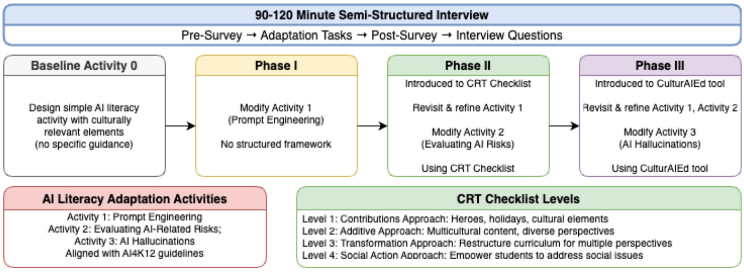
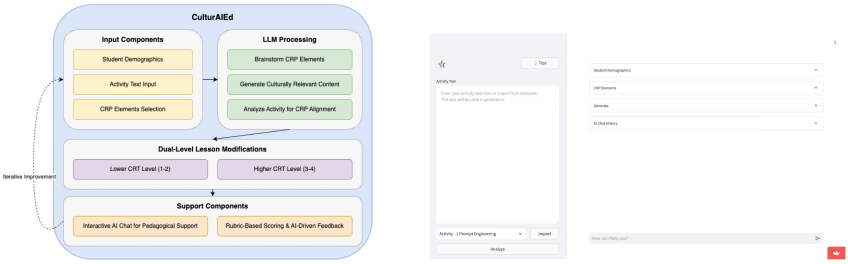


Fig. 1. Study Design

3.1 CulturAIEd Tool Design

CulturAIEd integrates principles from CRP/CRT to facilitate inclusive and effective lesson planning. Central to CulturAIEd's design is an LLM feature that provides culturally relevant content, such as examples, analogies, or contextual details, tailored specifically to the demographics of a teachers classroom.

This function supports educators in overcoming common barriers to CRT implementation by providing concrete, relatable connections that authentically reflect students' cultural backgrounds and lived experiences, thus fostering greater engagement and deeper academic understanding. Additionally, CulturAIEd provides two tiers of lesson modifications, basic and advanced, based on the CRT checklist [4], guiding teachers in culturally responsive adaptation. On the one hand, it provides teachers with flexible options suited to various instructional contexts or constraints, on the other hand, by explicitly showcasing differences between tokenistic and meaningful cultural integration, it encourages educators to move beyond surface-level adjustments towards genuine and substantive curricular changes. Complementing this, CulturAIEd employs rubric-based scoring and AI-driven feedback to evaluate and enhance lesson quality systematically. Teachers receive formative feedback highlighting areas for improvement, thereby promoting continuous professional growth through explicit and actionable recommendations. Finally, an interactive LLM chatbot provides just-in-time pedagogical support, simulating instructional coaching to address teachers' immediate needs. By embedding such support within teachers' routine planning processes, CulturAIEd advances educators competencies in delivering culturally responsive instruction, ultimately enriching student learning experiences. Before the interviews, we built a demo with streamlit [3] and gpt-4o-mini-2024-07-18 model.



(a) Design (b) Demo

Fig. 2. CulturAIEd Design

3.2 Study Procedure

In this IRB (Institutional Review Board)-approved study, participants completed 90–120 min semi-structured 1-on-1 Zoom sessions involving a pre-survey, adaptation tasks, post-survey, and interviews. The tasks focused on adapting AI literacy activities, aligned with AI4K12 guidelines [1], to include CRP elements. Participants started with completing **Activity 0** (designing a basic AI literacy lesson without guidance) as a baseline to observe CRP integration strategies. Participants then engaged in a structured adaptation process across three phases. In Phase I, they independently adapted a prompt engineering activity (**Activity**

1) without any guidance. In Phase II, participants employed a structured CRT Checklist [4] to adapt the previously addressed prompt engineering activity along with an additional task on evaluating AI-related risks (**Activity 2**), progressing through defined CRT levels: Contributions, Additive, Transformation, and Social Action. In Phase III, participants utilized CulturAIEd to refine their prior adaptations and to modify a new activity focused on AI hallucinations (**Activity 3**), exploring deeper and more meaningful integration of CRP strategies. Participants had access to generative AI tools (e.g., ChatGPT) and internet resources throughout.

3.3 Participants and Data Analysis

We recruited three secondary teachers from North America (P1, P2, P4) via Prolific [2] and one from East Asia (P3), using quota sampling to ensure racial and ethnic diversity. Participants included P1 (White female, science teacher, grades 9–12, virtual public school), P2 (Black male, music education teacher, grades 9–12, public school), P3 (Asian female, information technology teacher, grade 9, public school), and P4 (White female, science teacher, grades 9–12, public school). All participants had at least 4 years of teaching experience and varied AI-related experience. Participants received compensation at $30/hour.

We employed a qualitative-dominant mixed-methods approach, prioritizing thematic analysis [7] of interview transcripts to identify key themes due to the small sample size ($N = 4$). Survey data were analyzed descriptively, focusing on individual shifts in perceptions and practices of CRP/CRT integration with and without LLM-based tools in AI literacy settings. Methodological triangulation ensured reliability. Given the exploratory nature of the pilot study, findings offer insights rather than generalizable conclusions.

4 Preliminary Findings

4.1 AI Literacy in Schools: Teachers' Concerns and Practices

Survey data revealed teachers' active use of AI tools, such as ChatGPT and Turnitin, for lesson planning, content creation, and student support. However, challenges remain in verifying AI-generated content accuracy, managing subscriptions, and addressing varied student comfort levels with technology. Participants show varied familiarity with AI tools, balancing concerns with exploring potential educational uses. P1, teaching in a virtual public school, expressed worries about students potentially misusing AI, asking, *"How can we stop students from cheating all the time?"* This highlights broader concerns around academic integrity with AI access. Yet, there is recognition of the importance of teaching responsible AI usage, as P1 noted the necessity of guiding students to use AI constructively. For example, P1 used ChatGPT to simplify complex scientific texts and observed students creating concept maps with AI. P4 used LLM to generate multiple assignment examples from specified criteria for student selection. Despite these potential benefits, accuracy remains a concern: P1 noted

occasional hallucinations in AI-generated explanations, and P4 highlighted the ongoing need to develop clearer AI usage guidelines and policies.

4.2 CRP in Schools: Resources, Support, and Challenges

Interviews revealed varied levels of support for CRP across school environments. P2 indicated that although CRP is not explicitly mandated, it *"is highly expected,"* with administrators observing and providing feedback. They also noted professional development opportunities, mentioning *"they brought in a speaker and author who has written books on culturally relevant pedagogy."* In contrast, P4 remarked, *"when I was in college I didn't hear anything about culturally responsive teaching. And that was 15 years ago,"* adding that current workplace discussions are *"minimally, but not like at length."* P1 emphasized minimal support in previous experiences, noting, *"in other schools that I have worked for, it was not required there really were not a whole lot of supports."* These highlighted significant variability in institutional support and available resources for CRP. Teachers also described challenges in effectively implementing CRP, notably the considerable time commitment required. P2 emphasized the difficulty of authentically knowing students beyond superficial traits, stating, *"it really takes sitting down, knowing my students not just simply on skin color, but also socioeconomic background. It's a challenge because it takes time to do."* P1 shared similar concerns, initially questioning, *"how would I have time to incorporate all of this information that is relevant to my students and their experiences?"* Authenticity emerged as another issue: P1 recounted experiences at schools where activities labeled as "culturally responsive" lacked genuine connection to students' lives. P4 acknowledged uncertainty in making meaningful connections, admitting, *"I can just make assumptions. I guess I do my best."* Pre-survey results align with these perspectives. Participants reported moderate confidence (3 to 4 on a 5-point scale) in identifying and implementing CRP strategies, citing limited time and uncertainty about best practices as primary challenges.

4.3 CulturAIEd: Strengths and Opportunities

Participants praised CulturAIEd for streamlining CRP, particularly through its integration of student demographics, content customization, rubric-based feedback. Therefore, using CulturAIEd was viewed as highly efficient. Post-surveys confirmed high satisfaction, with all participants reporting increased confidence in implementing CRP modifications. In specific, teachers' confidence in "Identifying opportunities for cultural responsiveness in learning activities" rose significantly from 3–5 in the pre-survey to all 5 in the post-survey. P1 noted, "*I had not taken the time to consider culturally responsive practices, and it's making me realize its not so difficult to incorporate.*" P4 similarly expressed that the experience provided a helpful framework for designing culturally responsive activities. Identified opportunities for improvement include workflow enhancement, UI refinements, export functionality, and multilingual support.

Table 1. Strength and Opportunities of CulturAIEd

	Aspect	Supporting Quotes
Strengths	CRP Integration: Facilitates culturally responsive pedagogy (CRP) with actionable insights.	*Extremely confident that CulturAIEd is very helpful an enlightening tool* (P2); *I do see myself using CulturAIEd helpful for concrete examples* (P1).
	Streamlined Demographics & Modifications: Simplifies integrating student demographics and activity adjustments.	*Demographics information was straightforward nice to see integrated* (P2); *Solved the problem of time and research needed for CRP* (P1).
	Rubric Feature: Provides structured feedback for culturally responsive design.	*The rubric broke down areas teachers might not consider* (P2); *Liked the grading feedback on all categories* (P4).
	Post-Survey Results: High confidence in CRP modifications; satisfaction with outputs.	All participants rated themselves 5/5 in post-surveys; 3/4 strongly agreed CulturAIEd saved time and inspired new ideas.
Opportunities	Workflow Enhancements: Requests for automation and reduced manual input.	*Automatically add information to a certain part* (P1); *Integrate content with minimal copy/paste* (P4).
	User Interface (UI) Improvements: Need for a more intuitive interface.	*The UI could use a little work* (P1).
	Exporting Capabilities: Demand for direct export to common formats (e.g., Google Docs).	*Export to Google Docs or Word would be helpful* (P1); *Output in a document format probably easiest* (P4).
	Multi-Language Support	*More explicit multi-language support* (P3).

5 Discussion and Future Work

Results indicated a notable positive shift in teacher perceptions and practices regarding CRP after using CulturAIEd. We framed the study as a preliminary exploration to gather qualitative feedback. Accordingly, the improvements in confidence and efficiency should be interpreted as promising trends rather than conclusive evidence of efficacy. Future larger-scale studies (with control comparisons) will be needed to test CulturAIEds impact more rigorously.

Despite these positive changes, it is crucial to acknowledge the risks that come with using LLMs in this context. LLMs learn from vast datasets that may contain biases or stereotypes. If not carefully mitigated, the model could inadvertently produce content that reinforces simplistic cultural clichs or even offensive stereotypes, thereby undermining the culturally relevant pedagogy we seek to advance. Existing literature also warns against the risk of "routinizing" CRP [18]. Tools driven by LLMs, like CulturAIEd, could potentially become viewed as complete solutions rather than foundational resources. Although these tools support teaching through brainstorming, summarizing, and targeting, they cannot replace teachers' contextual understanding. Effective training is essential to help teachers adapt and refine AI suggestions for diverse student backgrounds.

Future research should examine the long-term impacts of LLM-supported tools like CulturAIEd on teacher practices and student outcomes in CRP across different subjects. Longitudinal studies could provide deeper insights into the sustained integration of CRP in lesson planning and classroom implementation. Moreover, the current sample size of four participants is too small, and further data collection from a broader range of contexts is essential to enhance the reliability and generalizability of the findings. Improvements to CulturAIEd, such as an enhanced user interface, better export options, and multi-language capabilities, should be prioritized to serve diverse educators effectively. Lastly, integrating such tools within professional development programs to encourage teacher reflection and refinement beyond template-based outputs represents another promising research direction.

References

1. AI4K12. https://ai4k12.org/
2. Prolific. https://www.prolific.com/
3. Streamlit. https://streamlit.io/
4. Culturally Responsive Teaching Checklist. https://reimaginingmigration.org/resource-items/cultural-responsive-teaching-checklist/ (2019)
5. Aronson, B., Laughter, J.: The theory and practice of culturally relevant education: a synthesis of research across content areas **86**(1), 163–206 (2016). https://doi.org/10.3102/0034654315582066
6. Baytak, A.: The content analysis of the lesson plans created by ChatGPT and Google Gemini. Res. Soc. Sci. Technol. **9**(1), 329–350 (2024)
7. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
8. Byrd, C.M.: Does culturally relevant teaching work? An examination from student perspectives **6**(3) (2016). https://doi.org/10.1177/2158244016660744
9. Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., Barro, S.: AI Literacy in K-12: a systematic literature review. IJ STEM Ed **10**, 29 (2023). https://doi.org/10.1186/s40594-023-00418-7
10. Dee, T., Penner, E.: The causal effects of cultural relevance: evidence from an ethnic studies curriculum **54**(1), 127 (2017)

11. Druga, S., Otero, N., Ko, A.J.: The Landscape of Teaching Resources for AI Education. In: Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education, vol. 1, pp. 96–102. ACM (2022). https://doi.org/10.1145/3502718.3524782
12. Eguchi, A., Okada, H., Muto, Y.: Contextualizing AI education for K-12 students to enhance their learning of AI literacy through culturally responsive approaches. KI - Künstliche Intelligenz **35**(2), 153–161 (2021). https://doi.org/10.1007/s13218-021-00737-3
13. Gay, G.: Culturally Responsive Teaching: Theory, Research, and Practice. Teachers College Press (2000)
14. Hammond, Z., Jackson, Y.: Culturally Responsive Teaching and the Brain (2015)
15. Johnson, A., Elliott, S.: Culturally relevant pedagogy: a model to guide cultural transformation in STEM departments **21**(1), 21.1.35 (2020). https://doi.org/10.1128/jmbe.v21i1.2097
16. Laak, K.J., Aru, J.: Generative AI in K-12: opportunities for learning and utility for teachers. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, pp. 502–509. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64315-6_49
17. Ladson-Billings, G.: Toward a theory of culturally relevant pedagogy **32**(3), 465–491 (1995). https://doi.org/10.3102/00028312032003465
18. Leonard, J., Napp, C., Adeleke, S.: The complexities of culturally relevant pedagogy: a case study of two secondary mathematics teachers and their ESOL students **93**(1), 3–22 (2009). https://doi.org/10.1353/hsj.0.0038
19. Li, H., Xiao, R., Nieu, H., Tseng, Y.J., Liao, G.: "From unseen needs to classroom solutions": Exploring AI literacy challenges & opportunities with project-based learning toolkit in K-12 education. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 29145–29152 (2025)
20. Long, D., Magerko, B.: What is AI Literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16. ACM (2020). https://doi.org/10.1145/3313831.3376727
21. Mette, I.M., Nieuwenhuizen, L., Hvidston, D.J.: Teachers' perceptions of culturally responsive pedagogy and the impact on leadership preparation: lessons for future reform efforts. Int. J. Educ. Leadership Preparation **11**(1), n1 (2016)
22. Moorhouse, B.L., Ho, T.Y., Wu, C., Wan, Y.: Pre-service language teachers' task-specific large language model prompting practices. RELC J., 00336882251313701 (2025). https://doi.org/10.1177/00336882251313701
23. Roy, P., Staunton, R., Poet, H.: ChatGPT in lesson preparation - a teacher choices trial. https://doi.org/10.1186/ISRCTN13420346
24. Samuels, A.J.: Exploring culturally responsive pedagogy: teachers' perspectives on fostering equitable and inclusive classrooms. Srate J. **27**(1), 22–30 (2018)
25. Tseng, Y.J., et al.: ActiveAI: the effectiveness of an interactive tutoring system in developing K-12 AI literacy. In: European Conference on Technology Enhanced Learning, pp. 452–467. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72315-5_31

Comparison of Explainable Machine Learning Methods for Early Prediction of Student Performance in Programming Courses

Ziwei Wang[1(✉)], Irena Koprinska[1], María Martínez-Ballesteros[2], Alicia Troncoso[3], and Bryn Jeffries[1]

[1] School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
zwan0684@uni.sydney.edu.au, {irena.koprinska,bryn.jeffries}@sydney.edu.au
[2] Department of Computer Languages and Systems, University of Seville, 41012 Seville, Spain
mariamartinez@us.es
[3] Data Science and Big Data Lab, Pablo de Olavide University, 41013 Seville, Spain
atrolor@upo.es

Abstract. We compare the performance of intrinsically explainable methods (decision trees and explainable boosting machines) with a black-box method (neural networks) with post-hoc explanations provided by SHAP, LIME and association rules, for the task of early prediction of student performance in a large-scale programming course. All three prediction methods obtained high accuracy, and all explainers were able to provide actionable insights by highlighting important features and key tasks students should complete. Decision trees provided compact rules, which were complemented by the more detailed behavioral patterns identified by the association rules. We observed high similarities between the feature ranking methods. We discuss the strengths of the methods in identifying key factors affecting student performance and providing explanations.

1 Introduction

Early prediction of students' final course performance is important for both educators and students. It can help educators to identify at-risk students and provide targeted support, and students to adjust their behavior to achieve better learning outcomes. The prediction methods need to be accurate but also explainable, to ensure trust, fairness and deployment in practice.

Decision Trees (DTs) are traditional intrinsically explainable (glass-box) methods, and have been widely used for predicting student performance [1,3,6,7,15,17]. Explainable Boosting Machines (EBMs) [8] are a newer state-of-the-art method combining DTs with gradient boosting. They are considered intrinsically explainable as they provide an in-built feature importance ranking.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 161–168, 2025.
https://doi.org/10.1007/978-3-031-99264-3_20

Neural Networks (NNs), on the other hand, are black-box methods. They have been very popular in the last few years due to the advances in deep learning, and have been applied in education settings, e.g. for modeling student learning [12] and dropout prediction [16]. Post-hoc explainability methods such as SHAP can be applied to the trained NNs to provide insights on how the decisions are made.

Research on explainable methods in education is still limited, and especially lacks comparison between glass-box methods and black-box with post-hoc explanations. Prior work focuses on them separately, e.g. on EBMs, demonstrating their applicability [2,5] or on NNs with post-hoc methods [14]. In this paper, our goal is to compare these two approaches.

The contributions of our work are: 1) We compare two intrinsically explainable models (DT and EBM) with a black-box method - NNs, used in conjunction with three post-hoc methods for providing explanations: SHAP, LIME and Association Rules (AR), 2) We demonstrate the application of ARs and highlight their potential; ARs haven't been applied as explainers of black-box models in education. 3) We investigate the performance of these methods in a large-scale K12 programming course, unlike previous work which considers university courses and smaller cohorts. Ultimately, we aim to gain a deeper understanding of the strengths of the different methods in identifying the key factors that affect student performance, and in providing explanations.

2 Data

We use data from a Beginners stream of an online Python programming course for high school students. The data was provided in a fully anonymised format. The course spans five weeks, with two modules per week. Each module includes content slides with explanations and code examples, and programming tasks (problems) to submit for automarking. There are four programming tasks in each module, worth 10 marks each. The three grade categories for the student total course mark are: Poor [0–25%], Good (25–75%] and Excellent (75–100%].

There were 11,416 enrolled students; 5,743 of them attempted at least one problem in the first half of the course and were included in the analysis. Among them, 2,786 were classified as Poor, 1,755 as Good, and 1,202 as Excellent.

3 Methodology

We investigate if it is possible to predict the final course grade (Poor, Good and Excellent), based on student log data from the first half of the course - from week 1 to week 3 module 1 (i.e. from w1p1 to w3p1).

3.1 Feature Extraction and Selection

We extracted features characterising student activity from each content slide, problem and module, as shown in Table 1. More information about the features

can be found in [15]. To select a smaller subset of informative features, we applied the correlation-based feature selection algorithm [4] which resulted in 19 features from the 6 highlighted categories in Table 1; these features were used to train DT, EBM and NN classifiers.

Table 1. Extracted features from content slides, problems and modules, those selected by CFS are in dark black. `wipj` - module week i part j, `ql` - problem l, `sk` - slide k.

Feature Name	Description
	Content Slides
w*ipj*-s*k*	Completion of slide (completed/not completed)
	Problems
w*ipj*-q*l*-runs	Count of times student run their code for the problem
w*ipj*-q*l*-subs	Count of times student submitted code for marking
w*ipj*-q*l*-sessions-outer	Count of sessions between completion of previous and current problem Session: sequence of events with no more than 60 seconds between them
w*ipj*-q*l*-sessions-inner	Count of sessions from first attempt to completion of current problem
w*ipj*-q*l*-active-time-outer	Total session time between completion of previous and current problem
w*ipj*-q*l*-active-time-inner	Total session time from first attempt to completion of current problem
w*ipj*-q*l*-overall-time-outer	Spanned time between completion of previous and current problem
w*ipj*-q*l*-overall-time-inner	Spanned time from first attempt to completion of current problem
w*ipj*-q*l*-first-attempt	Days from release until the student's first attempt
	Modules
w*ipj*-runs	Count of terminal runs within this module
w*ipj*-subs	Count of submits within this module
w*ipj*-slides-completed-pc	Percentage of content slides completed
w*ipj*-sessions	Count of sessions within this module
w*ipj*-active-time	Total active time within module
w*ipj*-overall-time	Overall time within module (span of all sessions)
w*ipj*-longest-session	Duration of the longest session
w*ipj*-mean-session-time	Average duration of the sessions
w*ipj*-time-efficiency	Total active time/total time

3.2 Intrinsically Explainable Methods: DT and EBM

DTs are one of the most prominent examples of inherently interpretabable methods. The tree can be converted into a set of if-then rules which are easy to visualise and understand by humans. We employed the REPTree version of DTs, which uses reduced error pruning to generate compact trees.

EBMs are a different type of inherently interpretable methods. They are an example of generalized additive models, combining decision trees using gradient boosting. EBMs generate feature importance scores reflecting the contribution of each feature to the prediction, enhancing comprehensibility for human users.

3.3 Post-hoc Explainable Methods: NN+SHAP, NN+LIME, NN+AR

NNs can capture complex patterns and typically offer strong predictive performance. However, they are black-box models and require post-hoc explainers to understand what has been learned. We built a three-layer feedforward NN and used grid search for hyperparameter tuning. After training, we applied three post-hoc methods to interpret NN's predictions: SHAP, LIME and AR.

SHAP [9] assigns a contribution score (Shapley value) to each feature. By considering all possible feature combinations, it provides a theoretically grounded approach to quantify each feature's impact on the model's prediction. **LIME** [13] also assigns a contribution score to each feature for each instance. It generates perturbed instances and trains a local linear model. **ARs** identify frequently co-occurring feature-value patterns and express them as if-then rules. To extract ARs we applied the MOQAR algorithm [10], which is based on an evolutionary process to learn the appropriate intervals for each feature.

4 Results and Discussion

4.1 Overall Performance

We trained and evaluated DT, EBM and NN to predict the final course performance using log data from the first half of the course. To evaluate performance, we used 70% of the data as training and 30% as test set.

All three models obtained high accuracy: DT: 85.4%, EBM: 87.2% and NN: 87.0%, with EBM and NN performing slightly better. DT was the fastest to train and required minimal hyperparameter tuning. EBM had the longest training time but was still fast on our large dataset and required little hyperparameter tuning. NN was relatively fast to train but required extensive hyperparameter tuning. These results indicate that traditional methods such as DTs are highly competitive with newer methods such as EBMs and NNs trained with advanced optimisation techniques. Both EBM and DT are easy to use in practice by non-experts and offer high accuracy and interpretability.

4.2 Rule-Based Explanations

DTs. The generated DT (Fig. 1) is highly compact and interpretable - it contains only three rules and one feature - `first-attempt` for specific problems in weeks 2 and 3: `w2p1-q4` and `w3p1-q4`. This feature corresponds to the number of days from the release of the problem until the student's first attempt. It shows the importance of starting early to achieve good grades. The threshold ≥ 28 for class "Poor" indicate that the students did not complete this problem as material released at the start of week 2 needs to be completed in the remaining 4 weeks. The low misclassification rate further suggests that students who skip w2p1-q4 rarely do well on subsequent questions. Hence, early engagement with the identified key problems is a strong predictor of the final student grade.

```
w2p1-q4-first-attempt < 28
|   w3p1-q4-first-attempt < 7 : Excellent (333/85)
|   w3p1-q4-first-attempt >= 7 : Good (555/138)
w2p1-q4-first-attempt >= 28 : Poor (835/29)
```

Fig. 1. Generated DT. The leaf nodes show the number of examples in each branch and how many of them were misclassified.

Table 2. Examples of ARs for each class

Rules	Confidence	Accuracy	Coverage	Lift
w2p1-q4-first-attempt∈[28, 50] ∧ w3p1-q2-first-attempt∈[32, 50] ∧ w3p1-q3-active-time-inner∈[0, 318] ⇒ Poor	98.4%	97.6%	96.5%	2.051
w2p1-q3-first-attempt∈[0, 23] ∧ w2p1-q4-subs∈[1, 9] ∧ w3p1-q4-first-attempt∈[17, 50] ⇒ Good	92.8%	90.9%	69.9%	3.634
w2p1-q3-overall-time-inner∈[0, 181914] ∧ w2p2-q1-active-time-inner∈[11, 1092] ∧ w3p1-q3-active-time-inner∈[0, 1246] ∧ w3p1-q4-first-attempt∈[0, 19] ⇒ Excellent	85.7%	94.5%	95%	3.236

ARs. ARs are effective in uncovering detailed and interpretable behavioral patterns that are often missed by global models such as DT. While ARs may generate overlapping or redundant rules, their strength lies in descriptive pattern discovery, offering rich insights into how students engage with course material.

In the ARs obtained by the MOQAR algorithm, we observed consistent behavioral trends across the three performance levels. Table 2 presents examples of learned ARs for each class. The AR metrics are described in [11]. The rules have high confidence and accuracy, meaning a high reliability and probability for correct prediction. The coverage reveals the proportion of students in each class satisfying the rule. The large lift values indicate that there is a strong positive relationship between the antecedent and the consequent of the rule, and therefore, the consequent is very likely to occur when the antecedent has occurred.

The rules from Table 2 show that Poor-performing students do not attempt some problems before the end of the course or start very late as indicated by the interval values of `first-attempt` for problems `w2p1-q4` and `w3p1-q2`, and have limited working time (`w3p1-q3-active-time-inner`), pointing to superficial engagement. In contrast, Good students start earlier and make more submissions at the beginning of the course (`w2p1-q3-first-attempt`, `w2p1-q4-subs`), but become less engaged in later stages as indicated by the delayed first submission or no submission for problem `w3p1-q4`. Excellent students demonstrate proactive and sustained behaviour - they start early and are engaged in the middle of the course (`first-attempt` on `w3p1-q4`) and have higher working time on problems (`w2p2-q1-active-time-inner`). Consistent with the DT model, `first-attempt` on `w2p1-q4` and `w3p1-q4` are identified as important features. The observed AR patterns in the dataset are frequently based on `first-attempt` time and number of submissions `subs`.

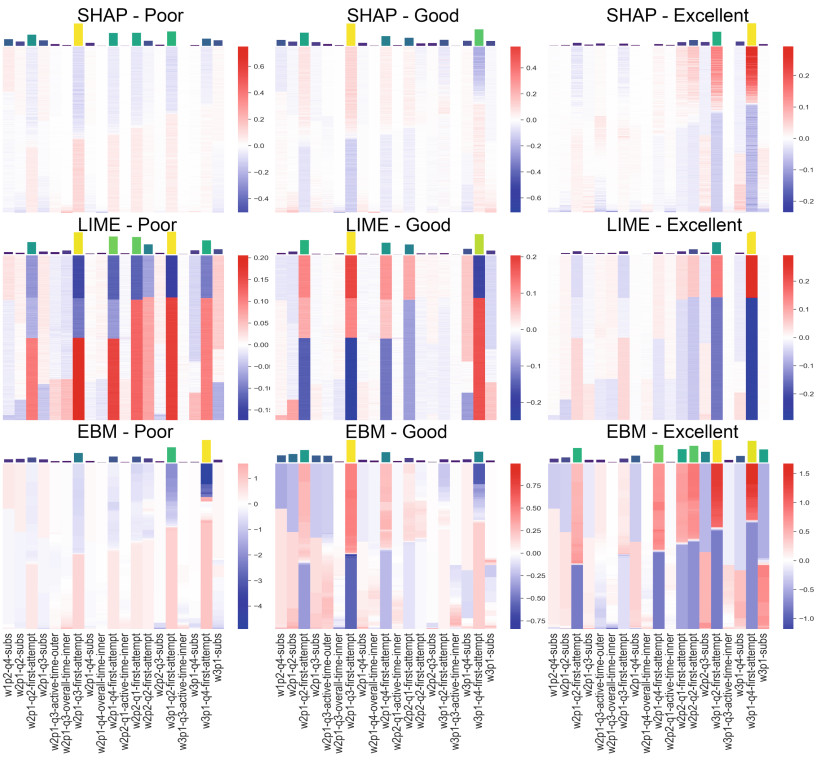


Fig. 2. Feature contribution scores for SHAP, LIME and EBM. Each column represents a feature, with red for positive and blue for negative contributions. Instances within each column are sorted from top to bottom by increasing raw feature values. The bar chart at the top of each heatmap represents the sum of absolute contribution scores. (Color figure online)

4.3 Feature Importance Explanations

SHAP, LIME and EBM explain the model by assigning a contribution score to each feature for every instance, which we visualise in Fig. 2 using heatmaps.

For SHAP, nearly all features with high scores are `first attempt` on specific tasks, underscoring the importance of early engagement with key programming tasks. For example, the most important feature for both Poor and Good is `first-attempt` on `w2p1-q3`. The high values of this feature contribute positively to class Poor and negatively to class Good, indicating that students who delay attempting this problem are more likely to achieve Poor grade. Consistent with the DT, `w3p1-q4-first-attempt` is the key feature distinguishing Excellent and Good students, highlighting again the importance of problem `w3p1-q4` which is the last problem in the first half of the course. Features based on submissions such as `w2p2-q3-subs` indicate that more submissions generally correlate with higher

performance, but there are cases of excessive submissions leading to negative contribution in predicting class Excellent, corresponding to struggling students.

LIME produces very similar results to SHAP. The top ranked features are very similar. For class Poor `w3p1-q2-first-attempt` is identified as almost equally important as the top ranked `w2p1-q3-first-attempt`. The contribution values show more noticeable changes between negative and positive influence, which may indicate critical points in the raw data that significantly impact predictions. Examining these changes may further enhance interpretability.

The most important features identified by EBM are similar to NN+SHAP and NN+LIME, although EMB has a very different underlying model than NNs - an ensemble of DTs with gradient boosting. Comparing EBM with NN+SHAP: for class Poor and Excellent, the top 2–3 features are the same for both, although the order is different, and for Good both the top three features and their order are the same. EBM assigns high scores to first attempt and number of submission for more problems compared to SHAP as strong indicators for predicting Excellent.

5 Conclusion

This study explored early prediction of student performance in a large-scale programming course, comparing the performance of intrinsically explainable methods (DT and EBM) with a black-box method (NN) with post-hoc explanations by SHAP, LIME, and AR. All three prediction methods obtained high accuracy, with EBM and NN performing slightly better than DT. All methods were able to provide actionable insights by highlighting important features such as starting early and identifying specific problems that students should complete. DT provided compact rules, AR - more detailed behavioral patterns complementing them. We observed high similarities between NN+SHAP, NN+LIME and EBM in the top ranked features. Ultimately, the choice of method depends on the specific application and required balance between accuracy and explainability. It may not be necessary to use complex black-box methods such as NNs, we recommend always comparing with glass-box methods such as DTs. Our study also demonstrated the potential of ARs as explainers for NNs.

Acknowledgment. The authors would like to thank the Spanish Ministry of Science and Innovation for the support within the projects PID2020-117954RB-C21, PID2020-117954RB-C22, PID2023-146037OB-C21, PID2023-146037OB-C22 funded by MICIU/AEI/ 10.13039/501100011033.

References

1. Asif, R., Merceron, A., Ali, S.A., Haider, N.G.: Analyzing undergraduate students' performance using educational data mining. Comput. Educ. **113**, 177–194 (2017)
2. Dsilva, V., Schleiss, J., Stober, S.: Trustworthy academic risk prediction with explainable boosting machines. In: International Conference on Artificial Intelligence in Education, pp. 463–475 (2023)

3. Gordon, C., Zhao, S., Vahid, F.: Ultra-lightweight early prediction of at-risk students in cs1. In: ACM Technical Symposium on Computer Science Education, pp. 764–770. ACM (2023)
4. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the International Conference on Machine Learning, pp. 359–366 (2000)
5. Jayasundara, S., Indika, A., Herath, D.: Interpretable student performance prediction using explainable boosting machine for multi-class classification. In: Proceedings of the International Conference on Advanced Research in Computing, pp. 391–396 (2022)
6. Koprinska, I., Stretton, J., Yacef, K.: Students at risk: detection and remediation. In: Proceedings of the International Conference on Educational Data Mining. pp. 512–515 (2015)
7. Liu, E., Koprinska, I., Yacef, K.: Early prediction of student performance in online programming courses. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 365–371 (2023)
8. Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 623–631 (2013)
9. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)
10. Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F., Riquelme, J.C.: Improving a multi-objective evolutionary algorithm to discover quantitative association rules. Knowl. Inf. Syst. **49**(2), 481–509 (2016)
11. Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F., Riquelme, J.C.: Obtaining optimal quality measures for quantitative association rules. Neurocomputing **176**, 36–47 (2016)
12. Piech, C., et al.: Deep knowledge tracing. In: Proceedings of the International Conference on Advances in Neural Information Processing Systems, vol. 28 (2015)
13. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: explaining the predictions of any classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
14. Swamy, V., Radmehr, B., Krco, N., Marras, M., Käser, T.: Evaluating the explainers: black-box explainable machine learning for student success prediction in moocs. In: Proceedings of the International Conference on Educational Data Mining, pp. 98–109 (2022)
15. Wang, Z., Koprinska, I., Jeffries, B.: Interpretable methods for early prediction of student performance in programming courses. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 115–123 (2024)
16. Xing, W., Du, D.: Dropout prediction in MOOCs: using deep learning for personalized intervention. J. Educ. Comput. Res. **57**(3), 547–570 (2019)
17. Zhang, V., Jefffries, B., Koprinska, I.: Predicting progress in a large-scale online programming course. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 810–816 (2023)

ScoreCLIQ: A Dynamic LLM-Based Framework for Item Difficulty Estimation

Soujatya Sarkar[1(✉)], Manikandan Ravikiran[1], and Rohit Saluja[1,2]

[1] Indian Institute of Technology Mandi, Mandi, India
s23106@students.iitmandi.ac.in
[2] BharatGen Consortium, Mumbai, India
https://github.com/souzatya/ScoreCLIQ

Abstract. Accurately estimating item difficulty is critical for building adaptive and fair educational assessments. However, most automated approaches treat difficulty as a static property, ignoring how subtle linguistic variations—such as stem rewording or distractor phrasing—can shift perceived complexity. We introduce ScoreCLIQ, a modular framework for dynamic item difficulty estimation that integrates model feedback with paraphrastic refinement. ScoreCLIQ first trains a BERT-based estimator on labeled difficulty scores. A Large Language Model (LLM) is then optimized via reinforcement learning to paraphrase items with high prediction error, using the frozen estimators output as a reward signal. Finally, the estimator is refined using a regularized objective that promotes prediction consistency between original and paraphrased items. This feedback-driven loop requires no additional human annotations and enables ScoreCLIQ to adapt to surface-level variation in item phrasing. On the BEA 2024 shared task dataset, ScoreCLIQ achieves a 12.46% reduction in RMSE over the previous state-of-the-art. Our results demonstrate that linguistic sensitivity in difficulty estimation can be learned through model-guided self-supervision.

Keywords: Item Difficulty Estimation · LLM · Reinforcement Learning

1 Introduction

Accurately estimating the difficulty of test items is central to building effective and fair assessments. Traditional psychometric approaches, such as Item Response Theory (IRT) [17], model item difficulty using learner response data, but require extensive pretesting and do not generalize easily to newly authored or modified items. To address this, recent work has leveraged natural language processing (NLP) techniques to predict item difficulty directly from the question text. Early efforts focused on handcrafted linguistic features [2], while more

S. Sarkar and M. Ravikiran—Equal Contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 169–176, 2025.
https://doi.org/10.1007/978-3-031-99264-3_21

recent approaches use pretrained transformer models such as BERT [4] to encode item content and regress scalar difficulty scores [5,9,18–21,23].

Despite their success, these models typically treat difficulty as a static property of the input, assuming that minor changes in phrasing or structure do not affect perceived difficulty. However, evidence from psychometrics and cognitive science suggests otherwise: small modifications in surface form—such as stem rewording, distractor phrasing, or option ordering—can significantly alter how difficult an item appears [3,10]. Nevertheless, existing models rarely account for such linguistic sensitivity. Even large language models (LLMs), though powerful for text generation, are commonly used in static prediction pipelines without adaptive feedback mechanisms [19].

Concurrently, other strands of NLP research have introduced techniques that are potentially relevant to item difficulty modeling. Reinforcement learning (RL) has been used to optimize generation objectives beyond token-level loss [1,16]. LLMs have been deployed as weak annotators or learned reward models [12,13], enabling self-supervised improvement. Data augmentation through paraphrasing has improved generalization in classification tasks [6,15]. Yet, these methods remain siloed from difficulty estimation, and prior work does not integrate paraphrasing, feedback-driven refinement, and difficulty prediction into a unified pipeline.

This paper introduces ScoreCLIQ, a framework for dynamic item difficulty estimation that explicitly accounts for linguistic variation. The framework consists of three stages: (i) a base estimator is trained to predict difficulty scores from question text; (ii) a large language model is fine-tuned using reinforcement learning to paraphrase items with high prediction error, where the frozen estimator serves as a weak reward signal; and (iii) the estimator is refined via a consistency-based regularization loss that aligns predictions across original and paraphrased versions. All components of ScoreCLIQ operate without additional human supervision, and the framework is applicable across item formats and domains.

We evaluate ScoreCLIQ on the BEA 2024 shared task dataset and observe a 12.46% reduction in RMSE over the previous state-of-the-art, demonstrating the effectiveness of feedback-driven paraphrasing for linguistic robustness in item difficulty modeling. The remainder of this paper is organized as follows: Sect. 2 introduces the ScoreCLIQ framework in detail. Section 3 presents the experimental setup and results. Section 4 concludes with key insights and future directions.

2 The ScoreCLIQ Framework

We propose ScoreCLIQ, a modular training framework for improving item difficulty estimation through feedback-guided paraphrasing. The framework consists of three components: (i) a baseline estimator trained on item-difficulty pairs, (ii) a large language model (LLM) fine-tuned via reinforcement learning to generate semantically equivalent paraphrases that align with difficulty scores, and (iii) a

regularized retraining step that encourages consistency across original and paraphrased inputs. Figure 1 illustrates the details of the proposed pipeline, which we discuss next.

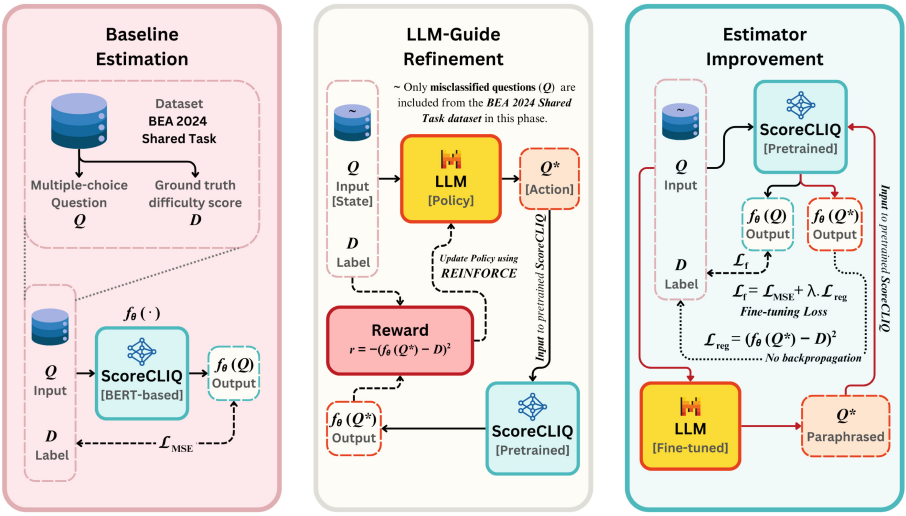


Fig. 1. The detailed three-stage training pipeline of ScoreCLIQ

Problem Setup. Let $\mathcal{D} = \{(Q_i, D_i)\}_{i=1}^{N}$ be a dataset of N items, where Q_i is a natural language question and $D_i \in \mathbb{R}$ is its human-annotated scalar difficulty score. Item difficulty refers to how challenging an item is for an average learner, typically derived from expert annotation or learner performance statistics. We aim to learn a predictor $\hat{D}_\theta : \mathcal{Q} \to \mathbb{R}$, parameterized by θ, that minimizes error on $\mathcal{D}$ while remaining robust to linguistic variation in Q_i.

Baseline Estimation. We initialize a base estimator f_θ using a BERT encoder followed by a regression head and train it using mean squared error:

$$\mathcal{L}_{\text{base}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (f_\theta(Q_i) - D_i)^2 \tag{1}$$

After training, we compute per-sample errors $\epsilon_i = |f_\theta(Q_i) - D_i|$ and define a high-error subset:

$$\widetilde{\mathcal{D}} = \{(Q_i, D_i) \in \mathcal{D} \mid \epsilon_i > t\}$$

where t is a tunable error threshold. These misclassified items serve as input to the next stage.

LLM-Guide Refinement. For each $(Q_i, D_i) \in \widetilde{\mathcal{D}}$, we fine-tune a pretrained LLM g_ϕ to generate a paraphrased variant Q_i^* such that $f_\theta(Q_i^*)$ is closer to D_i. The paraphraser is trained using reinforcement learning with a reward based on prediction error:

$$R(Q_i, Q_i^*) = -\left(f_\theta(Q_i^*) - D_i\right)^2 \tag{2}$$

We condition g_ϕ on the prompt-injected input sequence that includes Q_i, the model prediction $\hat{D}(Q_i)$, and the true label D_i. The model is optimized using the REINFORCE algorithm [22]:

$$\nabla_\phi \mathbb{E}_{Q_i^* \sim g_\phi(Q_i)}[R(Q_i, Q_i^*)] \approx \mathbb{E}_{Q_i^*}\left[R(Q_i, Q_i^*) \nabla_\phi \log g_\phi(Q_i^* \mid Q_i)\right] \tag{3}$$

Here, f_θ is frozen during paraphraser optimization and serves as a weak reward model. We adopt this policy gradient approach to optimize a non-differentiable objective without requiring labeled paraphrase examples (Fig. 2).

Question	Mistral-7B-Instruct-v0.3	
	Paraphrased before RL stage	Paraphrased after RL stage
A 52-year-old man comes to the physician with his wife because of a 1-year history of excessive daytime sleepiness. He does not think the symptoms are problematic, but his wife is concerned because he sometimes falls asleep on the sofa early in the evening when guests are present. He also once fell asleep while driving at night and drove off the road, narrowly avoiding injury. His wife says that he has always snored loudly, and over the past year, he has had episodes of choking or gasping for breath while sleeping. He is 178 cm (5 ft 10 in) tall and weighs 105 kg (231 lb); BMI is 33 kg/m2. His pulse is 76/min, respirations are 14/min, and blood pressure is 150/76 mm Hg. Physical and neurologic examinations show no other abnormalities. Which of the following is most likely to confirm the diagnosis? *True [0.18]*	Polysomnography (sleep study) would be most likely to confirm the diagnosis of obstructive sleep apnea (OSA). Given the symptoms of excessive daytime sleepiness, loud snoring, and episodes of choking or gasping for breath during sleep, as well as his BMI of 33 kg/m2, OSA is a strong possibility. A sleep study can objectively measure various parameters during sleep, including apnea-hypopnea index (AHI), oxygen desaturation, and arousals, which can help confirm the diagnosis of OSA. It is essential to note that the patient's age, sex, and weight are risk factors for OSA. OSA is a common sleep disorder characterized by repeated episodes of complete or partial upper airway obstruction during sleep, leading to disrupted sleep and daytime symptoms. In addition to the abovementioned symptoms, OSA can also cause hypertension, heart disease, and other complications if left untreated. It is crucial to discuss the risks and benefits of treatment options with the patient, including continuous positive airway pressure (CPAP) therapy, oral appliances, and surgery. Lifestyle modifications, such as weight loss, smoking cessation, and avoidance of alcohol and sedatives, can also help alleviate symptoms and improve quality of life. *Predicted [0.52]*	A 52-year-old man visits the physician with his wife due to persistent daytime drowsiness lasting one year. While the patient dismisses the symptoms as insignificant, his wife expresses concern over his tendency to doze off early in the evening when hosting guests, as well as an incident where he fell asleep while driving at night, nearly causing an accident. She reports a lifelong history of loud snoring and mentions that, over the past year, he has exhibited nighttime episodes of choking or gasping for air. The patient's height is 178 cm (5 ft 10 in), weight is 105 kg (231 lb; BMI 33 kg/m²), and vital signs include a pulse of 76/min, respirations of 14/min, and blood pressure of 150/76 mm Hg. Physical and neurologic examinations are otherwise unremarkable. Which diagnostic test is most likely to confirm the underlying condition? *Predicted [0.24]*

Fig. 2. A qualitative sample showcasing the difficulty-aware paraphrasing ability of the LLM-Guide after refinement stage. *Colour codes:* bad, medium, good, metadata

Estimator Improvement. The paraphrased items $\{Q_i^*\}$ are used to regularize the original estimator. We do not treat them as new samples; rather, they impose a consistency constraint:

$$\mathcal{L}_{\text{reg}}(\theta) = \frac{1}{|\widetilde{\mathcal{D}}|} \sum_{(Q_i, D_i) \in \widetilde{\mathcal{D}}} \left(f_\theta(Q_i^*) - D_i\right)^2 \tag{4}$$

The final loss for retraining becomes:

$$\mathcal{L}_{\text{final}}(\theta) = \mathcal{L}_{\text{base}}(\theta) + \lambda \mathcal{L}_{\text{reg}}(\theta) \tag{5}$$

where λ controls the regularization strength. This formulation encourages semantic stability: predictions remain consistent across paraphrastic transformations, improving generalization without additional annotations. We define a paraphrase in this setting as a transformation that maintains semantic equivalence while modifying surface features such as phrasing, vocabulary, or distractor wording.

3 Experiments, Results and Ablation Studies

We evaluate ScoreCLIQ on the BEA 2024 Shared Task dataset [23], which contains multiple-choice clinical questions annotated with scalar difficulty scores. Each item includes a question stem, answer options, the correct key, and the correct answer text. We concatenate these components into a single input Q_i, paired with its ground-truth difficulty $D(Q_i)$, forming the dataset $\mathcal{D}$.

We split the 466-question training set into 70% training and 30% validation subsets. The 201-question official test set is reserved for final evaluation. The base estimator f_θ is initialized using BERT [4]. For guided paraphrasing, we experiment with Mistral-7B-Instruct-v0.3 [14], Llama-3.2-1B-Instruct, and Llama-3.2-3B-Instruct [11] as the LLM g_ϕ. We selected BERT as the base estimator f_θ due to its widespread use in prior item difficulty estimation systems and its efficiency for low-resource fine-tuning.

For the paraphraser g_ϕ, we chose decoder-only models (Mistral-7B & LlaMA variants) due to their strong performance on instruction-following tasks and their compatibility with prompt-injected generation and token-level policy gradients. We leave the exploration of encoder-decoder LLMs for future work and expect our reinforcement learning loop to generalize across architectures. The paraphraser is conditioned on prompt-injected input, *'Original item: [Q]; Predicted difficulty: [value 1]; True difficulty: [value 2]; Rewrite the item to better align with the true difficulty.'*. We use reward normalization and entropy regularization to stabilize REINFORCE training. All experiments are run with 3 random seeds; variance in RMSE was below 0.002. Final hyperparameters are $\lambda = 0.99$ and $t = 0.04$.

3.1 Main Results

Table 1 benchmarks ScoreCLIQ against the top-performing systems from BEA 2024. Our best configuration, using Mistral-7B as the LLM-Guide, achieves an RMSE of 0.246, significantly outperforming the previous state-of-the-art (UnibucLLM, RMSE = 0.281). ScoreCLIQ with Llama variants also yields consistent gains. Compared to UnibucLLM which uses outputs from multiple LLMs, our single-LLM variant of ScoreCLIQ delivers stronger performance. This highlights the value of our dynamic, feedback-guided refinement over static augmentation strategies.

Table 1. Benchmarking ScoreCLIQ's performance (Base: BERT; LLM in parentheses) against state-of-the-art (SOTA) models.

Method	RMSE ↓
ScoreCLIQ (Mistral-7B-Instruct)	**0.246**
ScoreCLIQ (Llama-3.2-3B-Instruct)	0.259
ScoreCLIQ (Llama-3.2-1B-Instruct)	0.272
UnibucLLM [19]	0.281
EduTec [9]	0.299
UPN-ICC [5]	0.303
ITEC (RF) [20]	0.305
BC (FEAT+ROBERTA) [7]	0.305
Rishikesh [8]	0.310
Dummy Regressor [23]	0.311

3.2 Ablation Studies

Table 2 presents a sensitivity analysis of the threshold t and regularization weight λ. For fixed $\lambda = 0.99$, performance improves as t is lowered from 0.1 to 0.04, allowing the model to focus on more difficult examples. However, reducing t further introduces noise. Similarly, tuning λ from 1.0 to 0.99 slightly improves results, balancing the influence of original and paraphrased samples. The model remains stable across a range of values.

Table 2. Benchmarking the effect of threshold t and regularization weight λ on ScoreCLIQ performance.

Config.	t	RMSE ↓
	0.10	0.291
	0.08	0.282
ScoreCLIQ	0.06	0.273
($\lambda = 0.99$)	**0.04**	**0.246**
	0.02	0.251

Config.	λ	RMSE ↓
	1.00	0.249
	0.99	**0.246**
ScoreCLIQ	**0.98**	**0.246**
($t = 0.04$)	0.97	0.247
	0.96	0.250

4 Conclusion

We introduced ScoreCLIQ, a dynamic item difficulty estimation framework that incorporates model-guided paraphrasing and consistency-based refinement. Our method improves prediction performance over static BERT baselines and leading systems from BEA 2024. With a best RMSE of 0.246, ScoreCLIQ achieves a

relative improvement of over 12% compared to the prior state-of-the-art. This result is obtained using a single LLM as a guide, compared to ensembles used in competitive systems.

The feedback-driven design of ScoreCLIQ enables linguistic robustness and demonstrates the promise of using LLMs as optimization agents rather than static generators. While computationally more intensive than baseline finetuning, the framework remains feasible on modern hardware (e.g., one A100 GPU) and requires no additional annotations. In future work, we aim to test ScoreCLIQ on other item formats and domains (e.g., cloze and short-answer tasks) to further assess its generalizability.

Acknowledgments. We thank the anonymous reviewers for their constructive feedback. This work is supported by and is part of BharatGen (https://bharatgen.tech/), an Indian Government-funded initiative focused on developing multimodal large language models for Indian languages, also funded by the Department of Science & Technology (DST).

References

1. Bahdanau, D., et al.: An actor-critic algorithm for sequence prediction. In: International Conference on Learning Representations (ICLR) (2017)
2. Beinborn, L., Zesch, T., Gurevych, I.: Predicting the difficulty of language proficiency tests. Trans. Assoc. Comput. Linguist. **2**, 517–530 (2014)
3. Bennett, R.E.: Validity and automated scoring: it's not only the scoring. Educ. Meas. Issues Pract. **18**(4), 9–16 (1999)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423/
5. Dueñas, G., Jimenez, S., Ferro, G.M.: Upn-icc at bea 2024 shared task: leveraging llms for multiple-choice questions difficulty prediction. In: Workshop on Innovative Use of NLP for Building Educational Applications (2024). https://api.semanticscholar.org/CorpusID:270766389
6. Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 567–573 (2017)
7. Felice, M., Duran Karaoz, Z.: The British council submission to the BEA 2024 shared task. In: Kochmar, E., et al. (eds.) Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pp. 503–511. Association for Computational Linguistics, Mexico City (2024). https://aclanthology.org/2024.bea-1.42/
8. Fulari, R., Rusert, J.: Utilizing machine learning to predict question difficulty and response time for enhanced test construction. In: Kochmar, E., et al. (eds.) Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pp. 528–533. Association for Computational Linguistics, Mexico City (2024). https://aclanthology.org/2024.bea-1.45/

9. Gombert, S., Menzel, L., Di Mitri, D., Drachsler, H.: Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (2024)
10. Gorin, J.S.: Cognition and assessment: a critical issue in the field of testing. Educ. Meas. Issues Pract. **26**(3), 21–29 (2007)
11. Grattafiori, A., et al.: The llama 3 herd of models (2024). https://doi.org/10.48550/ARXIV.2407.21783. arxiv:2407.21783
12. Grosse, R., et al.: Can large language models be used as reward models? In: NeurIPS 2023 (2023)
13. Huang, Y., Zhao, J., et al.: Self-instruct: aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560 (2023)
14. Jiang, A.Q., et al.: Mistral 7b (2023). https://doi.org/10.48550/ARXIV.2310.06825. arxiv:2310.06825
15. Kumar, D., Ray, S., Chakrabarti, S.: Data augmentation using pre-trained transformer models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2750–2764 (2020)
16. Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., Jurafsky, D.: Deep reinforcement learning for dialogue generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1192–1202 (2016)
17. Lord, F.M.: A Theory of Test Scores. Psychometric Society (1952)
18. Ram, G.V.R., Kesanam, A., M, A.K.: Leveraging physical and semantic features of text item for difficulty and response time prediction of usmle questions. In: Workshop on Innovative Use of NLP for Building Educational Applications (2024). https://api.semanticscholar.org/CorpusID:270766188
19. Rogoz, A.C., Ionescu, R.T.: Unibucllm: harnessing llms for automated prediction of item difficulty and response time. In: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (2024)
20. Tack, A., et al.: Itec at bea 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pp. 512–521 (2024)
21. Veeramani, H., Thapa, S., Shankar, N.B., Alwan, A.: Large language model-based pipeline for item difficulty and response time estimation for educational assessments. In: Workshop on Innovative Use of NLP for Building Educational Applications (2024). https://api.semanticscholar.org/CorpusID:270766172
22. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. In: Machine Learning, vol. 8, pp. 229–256. Springer, Heidelberg (1992). https://doi.org/10.1007/bf00992696
23. Yaneva, V., et al.: Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pp. 470–482. Association for Computational Linguistics, Mexico City (2024). https://aclanthology.org/2024.bea-1.39/



LLM Agents for Verifiable Question Generation and Grading

Jacob Levine(✉), Matt West, and Mariana Silva

University of Illinois, Urbana, IL, USA
{jlevine4,mwest,mfsilva}@illinois.edu

Abstract. Automated question-asking systems play a crucial role in enhancing student learning by enabling repeated practice and frequent testing with flexible scheduling. With the rapid advancement of large language models (LLMs), we propose a pipeline for generating questions using LLMs while ensuring reliability in high-stakes assessments. Rather than directly generating student-facing questions, our approach produces code for PrairieLearn, an autograding platform that supports extensive randomization and customizable components, allowing for individualized questions with immediate and accurate feedback. To improve question quality, we augment instructor prompts with relevant information retrieved from example PrairieLearn questions and documentation. Additionally, we implement custom validators to automatically detect and repair non-functional questions, reducing the need for manual debugging. We evaluate our pipeline through a user study and an ablation analysis, demonstrating that each component contributes to improved question functionality. Our approach significantly lowers the barrier for instructors unfamiliar with PrairieLearn, enabling them to benefit from automated question generation and grading.

Keywords: LLMs · Question Generation · Auto-graders

1 Introduction

Large language models (LLMs) have had an explosive rise over the past few years for their state-of-the-art performance on many traditional NLP tasks [1,8,10] and extreme versatility as a base for many useful tools [5,11,12,15,17,23]. One potential application in educational settings is automatic question generation and grading. Given course materials or instructor-selected concepts, these systems can generate relevant questions for students and automatically evaluate their responses without human intervention. They provide students with immediate feedback and generate different question versions, allowing them to practice until they achieve mastery [3]. Additionally, they enable courses to scale efficiently without extra grading staff.

A simple pipeline for automatic question generation involves using an LLM to generate questions based on an instructor's prompt, delivering them to students, and then leveraging LLMs to evaluate student responses. While LLMs

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 177–184, 2025.
https://doi.org/10.1007/978-3-031-99264-3_22

have been successfully used to grade certain types of student answers [24], their reliability remains insufficient for high-stakes assessments, where more robust evaluation guarantees are desired. Current LLMs typically assume a confident tone and will often rather make up information, while claiming complete confidence in their answers, rather than failing to return any answer. These false claims, often called hallucinations [13], make it difficult to determine whether an LLM's grading response accurately reflects the student's knowledge or results from a hallucination. Even LLMs trained to resist adversarial prompting techniques can still be vulnerable to jailbreaks [20]. Without additional verification, pure LLM grading pipelines are not well-suited for high-stakes assessments.

Instead, we propose a pipeline where an LLM provides instructions on *how* to programmatically generate and grade a class of questions, rather than doing the generating and grading directly. We implement this pipeline in an open-source question-asking platform (herein referred to as "PrairieLearn" to follow anonymity guidelines) that is able to generate and grade student questions based on predefined rules. PrairieLearn [21] is a highly customizable platform that allows for many different autograding setups, where "question generators" written in Python and HTML can generate randomized questions and autograde the resulting student answers [22]. This enables support for questions as diverse as asking students to design finite state machines that follow certain properties, requesting asymptotic bounds rather than direct answers, and Parson's programming problems [14], which require students to select and re-order lines of code in order to complete a certain task [16].

The rule-based generation and grading framework of PrairieLearn enables instructors to create reliable high-stakes exams. During an exam, all students receive similar questions generated according to predefined rules and are graded against the same objective standard. Additionally, the ability to generate numerous question variants supports asynchronous exams with immediate feedback, minimizing opportunities for cheating [19]. This approach also facilitates second-chance testing setups, which help reduce student anxiety and provide a more accurate assessment of their understanding [2].

However, PrairieLearn's backend has a steep learning curve. Creating questions requires knowledge of both HTML and Python, along with familiarity with PrairieLearn-specific components and libraries to handle various student input types and ensure reliable grading. As a result, many instructors who could benefit from PrairieLearn may struggle to generate the necessary code themselves. Our pipeline addresses this challenge by allowing users to describe the type of question they want to create and the desired evaluation criteria, leaving the LLM to generate the corresponding PrairieLearn code, as illustrated in Fig. 1.

2 Related Work

One common approach to achieving high performance on specific tasks is Retrieval-Augmented Generation (RAG) [6], which supplements a pre-trained LLM with an additional knowledge base. RAG models find documents from a knowledge base to add to the context window, usually by embedding both queries

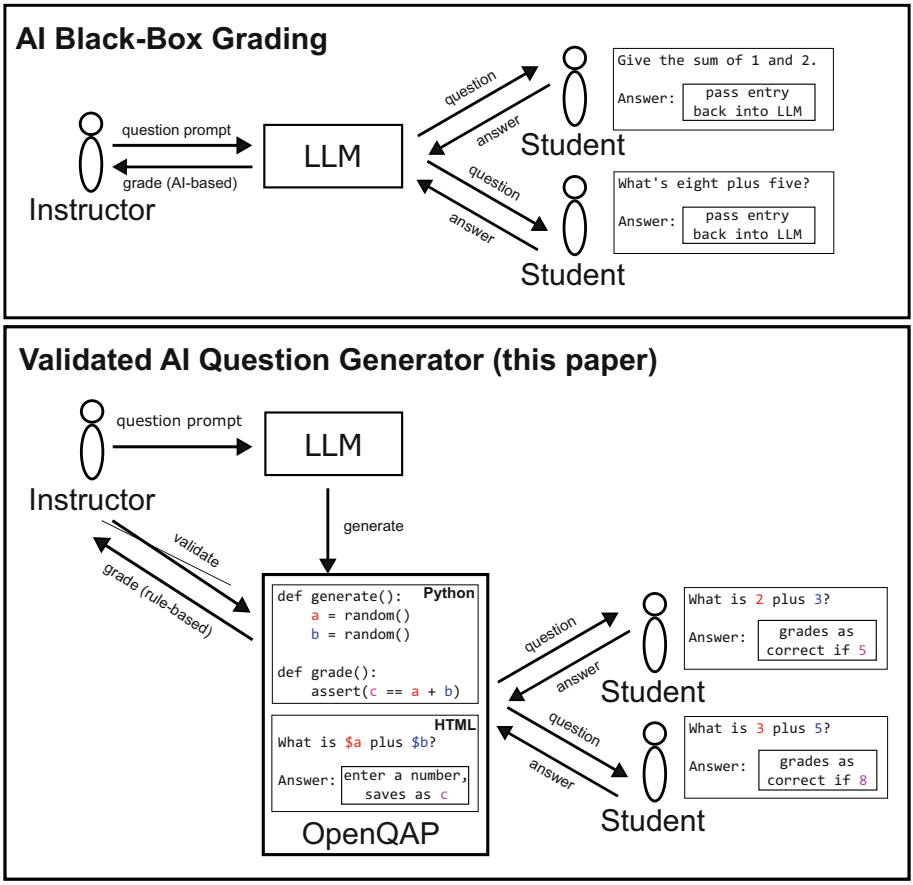


Fig. 1. AI Black Box Grading shows a simple end-to-end pipeline where questions are generated and evaluated solely based on LLM input. In contrast, our approach guides the LLM to define *how* the question should be created and graded, enabling instructors to review and verify that the grading process accurately assesses content knowledge.

and context into $\mathbb{R}^n$ using transformers and finding the documents closest to a query. However, more complex approaches, such as fine-tuning the embedding model [7] or by using polyencoders to rerank the top documents [18] can improve performance. This allows LLMs to utilize information outside of their internal knowledge to generate their responses, and therefore RAG approaches perform better on tasks that use knowledge-base information and have a lower hallucination rate [18].

Although RAG approaches can help reduce hallucination rates, additional validation steps may be necessary to ensure responses are accurate and useful. Developers of LLM pipelines may need to implement techniques to detect and modify incorrect responses. One common approach is automatic repair, where

LLMs are re-prompted to correct errors identified by external automated evaluation methods [4,9], an approach that we adopt.

3 Approach

We propose a pipeline to reliably generate PrairieLearn questions from user specifications using GPT-style LLMs. For each question, the user provides a specification that includes: (a) an overview of the question, including how internal parameters are generated, (b) the expected format for student responses, and (c) the criteria to determine correctness. Based on this input, our system generates a complete question, consisting of an HTML file with Mustache templates for layout using common PrairieLearn elements, as well as Python code for randomization and grading. The generated questions are returned to the user, allowing them to test the question and provide feedback for further refinements.

Prompt: Given a user's prompt, we format it into an LLM template, incorporating general guidelines on how to create questions in PrairieLearn, relevant PrairieLearn element documentation, and a few-shot examples of previously generated questions. We then prompt the LLM to generate a response containing the source code for one or more question files, following the structure demonstrated in the provided examples.

Document and Example Selection: For each generation or revision prompt, we retrieve documentation on one or more PrairieLearn elements, along with the source code of example questions that utilize those elements. To facilitate retrieval, files are embedded into $\mathbb{R}^{1586}$ using OpenAI's `text-embedding-3-small`. Given a query, we identify the five most relevant documents based on cosine similarity and use them as in-context examples, along with documentation related to how students should format their responses and specific documentation related to all OpenQAP elements used.

Validator: We perform a comprehensive set of syntactic and semantic checks, automatically re-prompting the LLM to correct any errors when necessary. To achieve this, we first parse the generated question file into an abstract syntax tree and traverse it to identify any tags that fall outside standard HTML or the subset of PrairieLearn elements we support. For each tag, we verify that all required properties are set, ensure that all defined properties match the documentation and have the correct types, and check for conflicts between properties. If discrepancies are found, our system re-prompts the LLM with specific feedback to correct the detected issues. Additionally, we execute all generated code and re-prompt with any errors.

User Interface: During the LLM question generation process, we save questions as "drafts" within the PrairieLearn interface. This allows users to preserve their progress and render questions for review during development while preventing them from inadvertently adding unfinished questions to an assignment or exam before they are finalized.

Upon entering the AI generation page, users are prompted to either continue working on an existing draft or create a new question. If they choose to create a new question, they are presented with text boxes for the three required aspects of the question, each accompanied by example inputs for guidance.

For LLM-based question revision, we provide a chat-style interface where users can request modifications. On the left, we display the complete prompt history, including the initial prompt and any human-generated revision requests, along with options to revert to previous versions or submit additional instructions. On the right, users can toggle between two views: a rendered question preview, where they can see how a student would interact with the question and test the grading process, and the generated source code, which allows for direct inspection of the implementation. Once satisfied with the modifications, users can finalize the question by assigning a question ID and title, making it available for use in assessments.

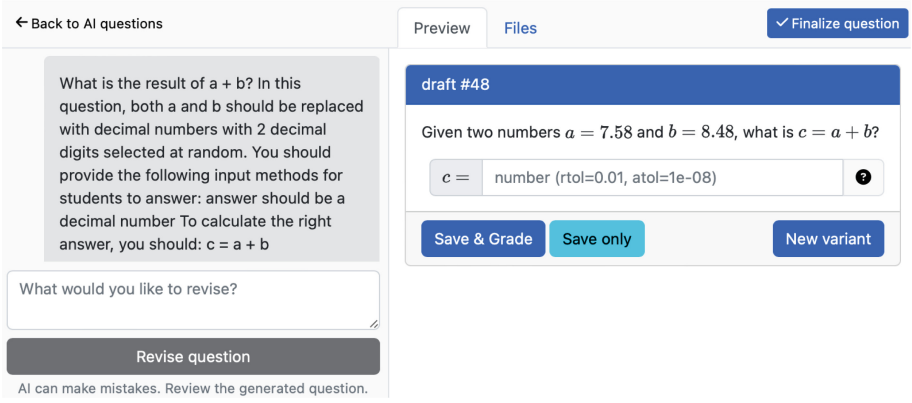


Fig. 2. User interface with an example question

User Studies: To evaluate the effectiveness of our pipeline in generating PrairieLearn questions, we conducted a small user study with six participants. Each participant was asked to create prompts for first-year undergraduate questions in physics, biology, and mathematics based on pictorial prompts, supplemented with examples of correct inputs and outputs. In total, we collected 24 question prompts and conducted an ablation study to assess the impact of different pipeline components on question quality. For each reduced pipeline configu-

ration, we generated one question per prompt (no user-initiated re-prompting) and evaluated it along two key axes:

Displays Correctly: Does the question render properly in PrairieLearn, allowing students to input and evaluate their answers?
Displays and Grades Correctly: Can someone with knowledge of the subject matter but no familiarity with PrairieLearns implementation answer the question successfully and have their response graded as correct?

Questions that meet the "Displays and Grades Correctly" criterion may still require minor visual adjustments, such as adjusting spacing or changing `^o(°)` to `\degree` (°). However, these adjustments should be manageable by users with no prior knowledge of PrairieLearn's setup, and the issues in unadjusted questions do not prevent users from obtaining and submitting a correct answer that is validated by the autograder.

4 Results

The results of the user study and ablations are shown in Table 1. The Full Pipeline performed well, with an 88% success rate of generating a correctly working and grading question directly from the user prompt (we did not test re-prompting by the users). The importance of pipeline components in order was: (1) the LLM model itself, with GPT-3.5 performing much worse than GPT-4o; (2) example questions RAG; (3) the validator with automatic re-prompting to fix errors; and (4) documentation RAG. This suggests that the example questions often provide implicit guidance not checked by the validator or detailed in the documentation.

Table 1. Success rates for generating PrairieLearn questions using our LLM pipeline on the 24 user-generated prompts from the user study. The first row (Full Pipeline) is the pipeline described in Sect. 3, while the later rows are ablations where one component at a time has been disabled.

Model	DisplaysCorrectly	Displays & Grades Correctly
Full Pipeline (using GPT-4o)	96%	88%
Downgrade to GPT 3.5	75%	38%
Remove Validator	92%	75%
Remove Example RAG	92%	71%
Remove Documentation RAG	83%	79%
Remove All RAG	63%	54%

5 Discussion and Conclusions

We introduce a pipeline for automatically generating and grading questions using LLMs. Rather than directly generating questions and grading student responses with LLMs, our approach focuses on generating PrairieLearn code which uses rule-based random generation and autograding of questions. This allows the PrairieLearn code to be validated, ensuring student knowledge is accurately tested, and is computationally cheaper than using the LLM to interact with each student individually.

Our system combines example and documentation retrieval and automatic repair to reliably generate PrairieLearn code, making it accessible to instructors who are not PrairieLearn experts. However, our current implementation is limited to a subset of PrairieLearn elements, and we have yet to conduct user studies on question revision through re-prompting. Moving forward, we plan to rectify these deficiencies and conduct further user studies.

References

1. Bang, Y., et al.: A Multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity (2023). https://doi.org/10.48550/arXiv.2302.04023. arXiv:2302.04023 [cs]
2. Emeka, C., Smith, D., Zilles, C., West, M., Herman, G., Bretl, T.: Determining the best policies for second-chance tests for STEM students. In: 2023 ASEE Annual Conference & Exposition Proceedings, p. 43019. ASEE Conferences, Baltimore (2023). https://doi.org/10.18260/1-2-43019
3. Guskey, T.R.: Closing achievement gaps: revisiting Benjamin S. Bloom's, "Learning for Mastery". J. Adv. Acad. **19**(1), 8–31 (2007). https://doi.org/10.4219/jaa-2007-704
4. Joshi, H., Cambronero, J., Gulwani, S., Le, V., Radicek, I., Verbruggen, G.: Repair is nearly generation: multilingual program repair with LLMs (2022). https://doi.org/10.48550/arXiv.2208.11640. arXiv:2208.11640
5. Kim, T.S., Choi, D., Choi, Y., Kim, J.: Stylette: styling the web with natural language. In: CHI Conference on Human Factors in Computing Systems, pp. 1–17. ACM, New Orleans (2022). https://doi.org/10.1145/3491102.3501931
6. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20. Curran Associates Inc., Red Hook (2020)
7. Li, J., Yuan, Y., Zhang, Z.: Enhancing LLM factual accuracy with RAG to counter hallucinations: a case study on domain-specific queries in private knowledge-bases (2024). https://doi.org/10.48550/arXiv.2403.10446. arXiv:2403.10446
8. Liang, P., et al.: Holistic evaluation of language models (2023). https://doi.org/10.48550/arXiv.2211.09110. arXiv:2211.09110
9. Liventsev, V., Grishina, A., Härmä, A., Moonen, L.: Fully autonomous programming with large language models. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1146–1155 (2023). https://doi.org/10.1145/3583131.3590481. arXiv:2304.10423
10. Lopez-Lira, A., Tang, Y.: Can ChatGPT forecast stock price movements? return predictability and large language models (2024). https://doi.org/10.48550/arXiv.2304.07619. arXiv:2304.07619

11. Lu, S., Duan, N., Han, H., Guo, D., Hwang, S.w., Svyatkovskiy, A.: ReACC: a retrieval-augmented code completion framework (2022). https://doi.org/10.48550/arXiv.2203.07722. arXiv:2203.07722
12. Luo, Z., et al.: WizardCoder: empowering code large language models with evol-instruct (2023). https://doi.org/10.48550/arXiv.2306.08568. arXiv:2306.08568
13. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906–1919. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.173
14. Parsons, D., Haden, P.: Parson's programming puzzles: a fun and effective learning tool for first programming courses. In: Proceedings of the 8th Australasian Conference on Computing Education, ACE '06, vol. 52, pp. 157–163. Australian Computer Society, Inc. (2006)
15. Parvez, M.R., Ahmad, W.U., Chakraborty, S., Ray, B., Chang, K.W.: Retrieval augmented code generation and summarization (2021). https://doi.org/10.48550/arXiv.2108.11601, arXiv:2108.11601
16. Poulsen, S., Gertner, Y., Cosman, B., West, M., Herman, G.L.: Efficiency of learning from proof blocks versus writing proofs. In: Proceedings of the 54th ACM Technical Symposium on Computer Science Education, SIGCSE 2023, vol. 1, pp. 472–478. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3545945.3569797
17. Rozière, B., et al.: Code llama: open foundation models for code (2024). https://doi.org/10.48550/arXiv.2308.12950. arXiv:2308.12950
18. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval Augmentation reduces hallucination in conversation. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3784–3803. Association for Computational Linguistics, Punta Cana (2021). https://doi.org/10.18653/v1/2021.findings-emnlp.320
19. Verma, A., Bretl, T., West, M., Zilles, C.: A quantitative analysis of when students choose to grade questions on computerized exams with multiple attempts. In: Proceedings of the Seventh ACM Conference on Learning @ Scale, pp. 329–332. ACM, Virtual Event (2020). https://doi.org/10.1145/3386527.3406740
20. Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: how does LLM safety training fail? In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems, vol. 36, pp. 80079–80110. Curran Associates, Inc. (2023). https://doi.org/10.5555/3666122.3669630
21. West, M., Herman, G., Zilles, C.: PrairieLearn: mastery-based online problem solving with adaptive scoring and recommendations driven by machine learning. In: 2015 ASEE Annual Conference and Exposition Proceedings, pp. 26.1238.1–26.1238.14. ASEE Conferences, Seattle (2015). https://doi.org/10.18260/p.24575
22. West, M., Walters, N., Silva, M., Bretl, T., Zilles, C.: Integrating diverse learning tools using the PrairieLearn Platform. Virtual Event (2021). https://cssplice.org/SIGCSE21/proc/SPLICE2021_SIGCSE_paper_10.pdf
23. Zhang, Q., Dong, J., Chen, H., Li, W., Huang, F., Huang, X.: Structure guided large language model for SQL generation (2024). https://doi.org/10.48550/arXiv.2402.13284. arXiv:2402.13284
24. Zhao, C., Silva, M., Poulsen, S.: Autograding mathematical induction proofs with natural language processing (2024). https://doi.org/10.48550/arXiv.2406.10268. arXiv:2406.10268

Computer Vision for Objects Used in Group Work: Challenges and Opportunities

Changsoo Jung(✉), Sheikh Mannan, Jack Fitzgerald, and Nathaniel Blanchard

Colorado State University, Fort Collins, CO 80523, USA
{Changsoo.Jung,sheikh.mannan,Jack.Fitzgerald,
Nathaniel.Blanchard}@colostate.edu

Abstract. Interactive and spatially aware technologies are transforming educational frameworks, particularly in K-12 settings where hands-on exploration fosters deeper conceptual understanding. However, during collaborative tasks, existing systems often lack the ability to accurately capture real-world interactions between students and physical objects. This issue could be addressed with automatic 6D pose estimation, i.e., estimation of an object's position and orientation in 3D space from RGB images or videos. For collaborative groups that interact with physical objects, 6D pose estimates allow AI systems to relate objects and entities. As part of this work, we introduce FiboSB, a novel and challenging 6D pose video dataset featuring groups of three participants solving an interactive task featuring small hand-held cubes and a weight scale. This setup poses unique challenges for 6D pose because groups are holistically recorded from a distance in order to capture all participants—this, coupled with the small size of the cubes, makes 6D pose estimation inherently non-trivial. We evaluated four state-of-the-art 6D pose estimation methods on FiboSB, exposing the limitations of current algorithms on collaborative group work. An error analysis of these methods reveals that the 6D pose methods' object detection modules fail. We address this by fine-tuning YOLO11-x for FiboSB, achieving an overall mAP_{50} of 0.898. The dataset, benchmark results, and analysis of YOLO11-x errors presented here lay the groundwork for leveraging the estimation of 6D poses in difficult collaborative contexts.

Keywords: 6D pose · collaborative group work · computer vision

1 Introduction

Recently, extensive research has shown the feasibility of an AI agent for collaborative groups in K-12 education [12,22]. However, these breakthroughs are typically driven by dialogic understanding [8,21] and largely ignore physical interactions across students and physical objects in these real-world settings.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 185–192, 2025.
https://doi.org/10.1007/978-3-031-99264-3_23

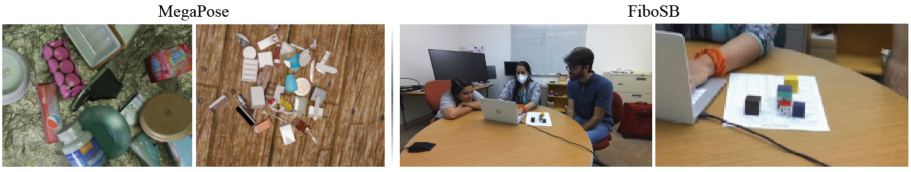


Fig. 1. Training image comparison between MegaPose and our FiboSB dataset. The left two images are from the synthetic training data of MegaPose [18], while the images on the right represent our FiboSB training data. The right-most image under FiboSB is zoomed in for illustration, and evaluations were conducted on the original images.

This gap is especially visible in collaborative activities that require hands-on manipulation, where solutions often rely on limited modalities such as verbal and textual input data rather than spatial and temporal clues. As a result, any AI system can only receive partial insight into group dynamics and task progress, limiting its ability to provide timely interventions and personalized support.

From an educational context, effective collaborative learning involves not only individual problem-solving but also group cooperation and physical interaction. Interactions during teamwork can include pointing, assembling, and manipulating objects to describe and prove concepts for elaborating deeper understanding between peers. Here, an AI-powered agent can play a crucial role; by monitoring how students handle and position objects, an agent can offer context-aware guidance as feedback and assessment during group work. Particularly in K-12 settings, where students vary in their developmental levels and learning styles, an agent that "sees" and "understands" these interactions can adapt instructions, prompt collaborative discussions, and highlight suggestions by each student's progress in the task.

Beyond facilitating collaboration, a precise awareness of physical space and object relationships is essential for understanding many concepts in education, especially in spatial reasoning. 6D pose estimation entails the ability to track the 3D position and orientation of objects[1], enabling detailed monitoring, such as how objects move or rotate and how they relate to each other in the environmental setting. This capability is particularly beneficial for younger or lower-grade students, who often depend on visual and tactile experiences to understand geometrical concepts: distance, shape, and scale. By recognizing real-time positions and orientations of objects, we can better measure students' performance, observe their decision-making processes, and provide elaborate feedback.

Educational studies have started incorporating tactile objects and 3D elements in lessons, benefiting students who lack spatial reasoning skills [1,23], providing a compelling opportunity to integrate 6D pose estimation into collaborative learning. In this paper, we investigate the potential of 6D pose estimation in supporting K-12 collaborative tasks and outline how this technique

[1] The term 6D comes from the need to predict the 3D translation and the 3D rotation of the object's pose.

can enhance spatial understanding, foster teamwork, and ultimately improve the overall learning experience.

To complement and extend the modalities studied in prior work [3,16,25], we introduce a novel 6D pose dataset called Fibonacci Small Blocks (FiboSB), which is adapted from the Weight Task Dataset (WTD) [15].

In summary, the key contributions of this study are as follows.

- Collection of a novel 6D pose dataset specifically designed for educational settings involving collaborative group tasks.
- Exploration of baseline performance of multiple state-of-the-art 6D pose methods on our dataset, highlighting the distinct challenges posed by collaborative group scenarios.

2 Dataset: Fibonacci Small Blocks (FiboSB)

To demonstrate the potential of using 6D pose estimation for collaborative group work in an educational setting, we introduce the FiboSB dataset. FiboSB is based on the Weights Task Dataset (WTD) [15], which involves a group of triads interacting with six colored blocks (two 1.5 $inch^3$ and four 2.0 $inch^3$) and a weight scale to determine the weights of each block. The group is initially given the weight of one block, and then they are tasked with finding the weights of the rest of the colored blocks, which follows a Fibonacci sequence. In FiboSB, we annotate the 6D poses of the colored blocks in the WTD so that we can train and evaluate 6D pose estimation models.

Predicaments in FiboSB. Detecting small blocks in the collaborative group work scene is a challenge (the two images on the right of Fig. 1). Occlusions among the blocks make the visibility of each block worse, as shown in the last column of Fig. 1, which frequently occur during the group task. The FiboSB dataset contains 25,381 annotated frames across 10 groups, with the number of frames per group ranging from 1,257 to 3,967. From the annotated frames, 133,263 object instances were annotated in total. On average, each frame contains 5.25 objects, indicating that multiple colored blocks frequently appear together. The blocks are often placed close to other blocks, leading to frequent occlusions that increase the complexity of the dataset. This obstacle leads to difficulties in estimating the exact object positions and their orientations. In addition, one pixel off on an annotation or prediction results in a huge error for small objects. For these reasons, recognizing the blocks and estimating their precise 6D pose predictions are critical requirements in the FiboSB dataset.

FiboSB vs. Other 6D Pose Datasets. To our knowledge, FiboSB is the first 6D pose dataset aimed at collaborative tasks in educational settings. Many 6D pose datasets [2,4,9] are targeted at various obstacles such as textureless and transparent objects [11,14]. Figure 1 shows examples of synthetic training

data for MegaPose; the images contain everyday objects, such as clocks, bottles, toys, etc., which have substantially different colors and shapes. This is in stark contrast to our dataset, where the blocks are objects used in a collaborative group task and are not the focus of the WTD.

3 Methodology

In this study, we explore state-of-the-art methods as baselines to evaluate the FiboSB dataset. Most 6D pose estimation methods use a two-stage pipeline: (1) object detection, which provides 2D bounding boxes for predicted objects, and (2) 6D pose estimation, which predicts the objects' 3D translations and orientations based on the interest regions from the first stage.

Object Detection Metrics: We evaluated the initial stage of the state-of-the-art 6D pose estimation approaches, which is the object detection modules, by employing the mAP_{50} metric (mean of Average Precision (AP) at Intersection of Union (IoU) threshold of 0.5) [6,7]. The values of mAP_{50} range between 0% to 100%, and a higher value represents better performance.

6D Pose Estimation Metrics: Since an object's position in 3D space is decided by its translation and rotation, the 6D pose estimation technique provides spatial information of the object. The object detection module delivers 2D spatial information for each object. Next, the 6D pose estimation module predicts corresponding translation and rotation in the 3D coordinate system.

For evaluation metrics of 6D pose estimation, we employed $Proj2D$ [10] and $ADD-S$ [26]. While $Proj2D$ assesses differences in 2D space, the $ADD-S$ metric quantifies errors in 3D space.

Experimental Details. To establish baseline performance on FiboSB, we train SOTA RGB-based 6D pose methods namely, CosyPose [17], RADet [19], and YOLOX-m-6D [20]. Furthermore, we evaluate MegaPose [18] to determine whether zero-shot-based approaches can provide reliable estimations for previously unseen objects in our collaborative setting. CosyPose, RADet, and YOLOX-m-6D are trained from scratch on our dataset using a group-wise split; groups 9 and 10 were assigned to the test set, and the remaining groups to the training set. The predictions are then assessed using the appropriate metrics as stated above.

4 Results

Although the SOTA methods show robustness on other (traditional) 6D pose datasets, the baselines were unreliable on our collaborative setting for both object detection and 6D pose estimation modules. We trace the reason for failure and then address the object detection issues with DETR and YOLO11-x [5,13].

Initial Evaluation of 6D Pose Estimation. We discovered that all of the models, except for MegaPose, failed to make any predictions during evaluation. Megapose received a poor score on the *ADD-S* metric (0.16) with a threshold of 0.1 diameter, which indicates that it struggled to make fine-grained predictions of the colored blocks using ground-truth bounding box information (Table 1); MegaPose has the most trouble at estimating pose for the yellow blocks with an average error in 3D distance of 157.53 mm from the ground truth labels.

Why is 6D Pose Failing? After obtaining the results from the evaluation of the 6D pose models, we were left with the burning question: what is causing CosyPose, RADet, and YOLOX-m-6D to not make any predictions? We decided to take a step back and analyze the multi-stage architectures of these models from the beginning, starting with the initial stage that performs object detection.

We first evaluated if there were any issues with our 6D pose implementations. Labbé et al. demonstrated that their method, MegaPose, achieves robust pose estimation performance, with average recall scores of 90.5 and 88.9 for ADD (0.1d) and Proj2D (5px) respectively, outperforming the Multi-Path method on the ModelNet dataset (RGB) [18]. It provides a good contrast to the results on FiboSB, suggesting that the issue was related to the increased difficulty of the FiboSB dataset.

Table 1. Overall metrics for 6D pose estimation module of MegaPose.

Metric	Red	Yellow	Green	Blue	Purple	Brown	Overall
3D Distance (mm) ↓	105.98	157.53	106.86	87.00	86.24	73.89	**106.17**
Proj2D (px) ↓	18.88	23.27	25.58	18.83	24.37	21.52	**22.11**
ADD-S (0.1d) ↑	0.17	0.12	0.08	0.17	0.15	0.42	**0.16**

Next, we performed an ablation of the 6D pose methods. We realized that the object detection modules were failing, i.e., they did not detect any of the blocks in the test set and thus caused the 6D pose stage of the models to make no predictions whatsoever. We evaluated the object detection modules individually and found that they received astonishingly low mAP_{50} scores–CosyPose with 0.004, RADet with 0.000, and YOLOX-m-6D with 0.005.

Addressing and Analyzing Object Detection Problem. After pinpointing the failure of the 6D pose estimation models as originating from the object detection modules, we explored more sophisticated object detection models such as DETR and YOLO11-x [5,13] to verify our FiboSB dataset. The DETR and YOLO11-x methods achieved 0.706 and 0.898 on the mAP_{50} metric respectively. Notably, the DETR model trained from scratch also failed on our dataset like the baselines. Table 2 shows the results of YOLO11-x using the mAP_{50} metric. We

assumed that YOLO11-x performed better because of additional data augmentations and multi-scale techniques. This experiment further proved to us that a more sophisticated object detection model is capable of accurately detecting the colored blocks.

Table 2. The performances of fine-tuned YOLO11-x with the mAP_{50} metric under our additional data. We assigned a larger portion of groups to the validation and test sets compared to the initial experiments (Sect. 3). Specifically, groups 1 to 4, groups 5 to 7, and groups 8 to 10 were distributed to train, validation, and test sets respectively. The first three rows compare the performance across groups in the test set. The last rows represent the overall results on the test set.

Additional Data	Red	Yellow	Green	Blue	Purple	Brown	All
Group 8	0.995	0.995	0.995	0.841	0.995	0.991	0.969
Group 9	0.995	0.995	0.995	0.782	0.000	0.456	0.704
Group 10	0.887	0.992	0.905	0.893	0.896	0.993	0.928
FiboSB	0.961	**0.990**	0.967	0.851	0.765	0.852	0.898

5 Discussion and Conclusion

6D pose estimation provides essential spatial context between students and objects in collaborative settings for AI agents. The agents enable us to trace student performance and infer the reasons behind object movements for immediate and objective feedback on collaborative group tasks such as in [24]. These capacities foster teamwork and improve learning experiences, and instructors can focus on higher-level advising and personalized support while the agent handles assessments and tracking the group process. As an initial study exploring the advantages, we focused on a simplified problem, the Fibonacci weight task, to explore the performances of the current state-of-the-art methods on our dataset. Based on current SOTA 6D pose estimation models, our findings reveal that existing object detection modules within these models lack the capabilities to even detect small objects in our collaborative setting. Moreover, the 6D pose estimation module still contains gaps to accomplish the fine-grained predictions. New 6D pose foundation models need to be developed that are able to detect small objects and produce precise 6D pose estimations in educational settings, collaborative or otherwise, such that an AI agent uses object detections to understand the problem at hand and eventually provide optimal guidance.

Acknowledgment. This material is based in part upon work supported by the National Science Foundation (NSF) under subcontracts to Colorado State University on award DRL 2019805 (Institute for Student-AI Teaming), and by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program.

Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. All errors are the responsibility of the authors.

References

1. Amir, M.F., Fediyanto, N., Rudyanto, H.E., Nur Afifah, D.S., Tortop, H.S.: Elementary students' perceptions of 3dmetric: a cross-sectional study. Heliyon 6(6), e04052 (2020). https://doi.org/10.1016/j.heliyon.2020.e04052. https://www.sciencedirect.com/science/article/pii/S2405844020308963
2. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D object pose estimation using 3D object coordinates. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 536–551. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_35
3. Bradford, M., Khebour, I., Blanchard, N., Krishnaswamy, N.: Automatic detection of collaborative states in small groups using multimodal features. In: International Conference on Artificial Intelligence in Education, pp. 767–773. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-36272-9_69
4. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: towards common benchmarks for manipulation research. In: 2015 International Conference on Advanced Robotics (ICAR), pp. 510–517. IEEE (2015)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
6. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision **111**, 98–136 (2015)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**, 303–338 (2010)
8. Grenander, M., Belfer, R., Kochmar, E., Serban, I.V., St-Hilaire, F., Cheung, J.C.: Deep discourse analysis for generating personalized feedback in intelligent tutor systems. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 15534–15544 (2021)
9. Guo, A., et al.: Handal: a dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 11428–11435. IEEE (2023)
10. He, X., Sun, J., Wang, Y., Huang, D., Bao, H., Zhou, X.: Onepose++: keypoint-free one-shot object pose estimation without cad models. Adv. Neural. Inf. Process. Syst. **35**, 35103–35115 (2022)
11. Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-less: an rgb-d dataset for 6d pose estimation of texture-less objects. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 880–888. IEEE (2017)
12. Houde, S., et al.: Controlling ai agent participation in group conversations: a human-centered approach. In: ACM International Conference on Intelligent User Interfaces (2025)

13. Jocher, G., Qiu, J.: Ultralytics yolo11 (2024). https://github.com/ultralytics/ultralytics
14. Jung, H., et al.: Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22498–22508 (2024)
15. Khebour, I., et al.: When text and speech are not enough: a multimodal dataset of collaboration in a situated task. J. Open Human. Data **10**(1) (2024)
16. Khebour, I.K., et al.: Common ground tracking in multimodal dialogue. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 3587–3602. ELRA and ICCL, Torino (2024). https://aclanthology.org/2024.lrec-main.318/
17. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: CosyPose: consistent multi-view multi-object 6D pose estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12362, pp. 574–591. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_34
18. Labbé, Y., et al.: Megapose: 6d pose estimation of novel objects via render & compare. arXiv preprint arXiv:2212.06870 (2022)
19. Li, Y., Huang, Q., Pei, X., Jiao, L., Shang, R.: Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. Remote Sens. **12**(3), 389 (2020)
20. Maji, D., Nagori, S., Mathew, M., Poddar, D.: Yolo-6d-pose: enhancing yolo for single-stage monocular multi-object 6d pose estimation. In: 2024 International Conference on 3D Vision (3DV), pp. 1616–1625. IEEE (2024)
21. Park, M., Kim, S., Lee, S., Kwon, S., Kim, K.: Empowering personalized learning through a conversation-based tutoring system with student modeling. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–10 (2024)
22. Rızvı, M.: Investigating ai-powered tutoring systems that adapt to individual student needs, providing personalized guidance and assessments. Eurasia Proc. Educ. Social Sci. **31**, 67–73 (2023)
23. Unal, H., Jakubowski, E., Corey, D.: Differences in learning geometry among high and low spatial ability pre-service mathematics teachers. Int. J. Math. Educ. Sci. Technol. **40**(8), 997–1012 (2009)
24. VanderHoeven, H., et al.: Trace: real-time multimodal common ground tracking in situated collaborative dialogues. arXiv preprint arXiv:2503.09511 (2025)
25. Venkatesha, V., et al.: Propositional extraction from natural speech in small group collaborative tasks. In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 169–180 (2024)
26. Zhang, Q., Xue, C., Qin, J., Duan, J., Zhou, Y.: 6d pose estimation of industrial parts based on point cloud geometric information prediction for robotic grasping. Entropy **26**(12), 1022 (2024)

From Sketch to Understanding: Exploring LLM-Based Assessment of Student-Drawn Science Models

Andy Smith[1](✉), Natalie Brezack[2], Wynnie Chan[2], Abby Lavine[2], Bradford Mott[1], Cathy Ringstaff[2], Mingyu Feng[2], and James Lester[1]

[1] Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA
{pmsmith4,bwmott,lester}@ncsu.edu
[2] WestEd, San Francisco, CA 94107, USA
{nbrezac,wchan,alavine,cringst,mfeng}@wested.org

Abstract. Drawing holds significant promise for supporting students' science learning. While drawing science models enables students to engage deeply with science content, teachers struggle to assess students' science sketches because it is prohibitively labor-intensive. Recent advances in generative AI models introduce the opportunity to automatically assess students' science drawings. We present a multimodal LLM-based science sketch assessment framework for automatically analyzing student-drawn science models. First, we collected a dataset of science models sketched by 8–12 year olds. Next, using evidence-centered design, a science model sketch rubric was developed, and the science models were assessed by human experts. Finally, our science sketch assessment framework was utilized to analyze the science models. Empirical evaluations of the assessment framework using two foundational multimodal models were performed. Findings from the evaluation indicate the science sketch assessment framework accurately assesses student science models, paving the way toward automated formative assessment for supporting students' sketch-based science modeling in the classroom.

Keywords: Sketch analytics · Multimodal LLMs · Science assessment

1 Introduction

Drawing encourages deep interaction with the subject matter and engages students in complex cognitive and learning processes, holding significant potential for enhancing student learning [1, 2]. An extensive meta-analysis of drawing-to-learn found that drawing yields significant improvements in learning relative to the status quo and produces transfer effects [3]. For elementary grade science instruction, U.S. standards include that students should be able to construct drawings or diagrams as representations of events or systems [4, 5].

Although learning-by-drawing offers considerable promise, there is growing literature on the importance of creating the proper conditions around learning-by-drawing to produce the strongest results. Drawing activities do not guarantee improved learning and,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 193–200, 2025.
https://doi.org/10.1007/978-3-031-99264-3_24

if not accompanied by appropriate support, can overwhelm students, increase cognitive load, and detract from learning [6, 7]. Adopting drawing activities in the classroom is also challenging because it places a significant burden on teachers to review and assess student drawings, especially in domains like science where teachers often lack specific training in the subject matter [8]. This paper presents an investigation of the effectiveness of multiple multimodal LLMs in evaluating student-drawn science models. Using the ScienceSketch sketch-based learning environment, we collected 62 models from 28 elementary students (aged 8–12 years). Using evidence-centered design (ECD), a rubric was developed and the student drawings were assessed by human raters. The drawings were then assessed by two multimodal LLMs using two different prompts. Findings indicate that the best performing LLM accurately assessed the student drawings, laying the foundation for automated sketch-based assessment for K-12 science education.

2 Related Work

Several systems have been used to assess and understand user-generated drawings. Multiple studies have used *Sketch Worksheets*, powered by the *CogSketch* system, which utilizes an analogical reasoning framework to interpret spatial relationships between glyphs [9–11]. The *SimSketch* system generated dynamic simulations based on elementary students' freehand drawings utilizing a system where students could annotate and label objects in the drawing space [12]. In the *Physics Playground* environment, students learn about physics by sketching simple machines such as levers, which they then manipulate [13–15]. *Physics Playground* creates motivating learning experiences that are enhanced with adaptive supports for learners. Sketch-based systems for undergraduate mechanical engineering have interpreted and provided feedback on truss and freebody diagrams [16]. More recent systems have utilized convolutional neural networks to assess student chemistry models [17] and third grade students' hand-drawn life science diagrams [18].

Generative AI systems, particularly LLMs, have seen a growing popularity in various automated assessment tasks, primarily natural language tasks. Multiple systems have utilized LLMs advanced text understanding capabilities to grade short-text answers in domains such as reading comprehension [19], K-12 science assessments [20], and third to sixth grade science assessments [21], often achieving results rivaling the previous state of the art. LLMs have also shown promise as effective tools for classifying dialogue acts between students [22, 23]. Our work builds on these systems but utilizes multimodal LLMs to assess student-drawn science models.

3 Methods

To explore the potential of learning-by-drawing, the ScienceSketch sketch-based learning environment was developed to support upper elementary students' science learning. ScienceSketch provides an online platform for students to draw models following the Next Generation Science Standards (NGSS), which can then be automatically scored to provide feedback. Currently, ScienceSketch focuses on fourth grade (ages 9–10) science standards that include science modeling around life sciences and physical sciences.

Student drawings were collected in ScienceSketch during a preliminary study to gather data for developing the scoring rubric and system. A researcher met with each student over video conference for one hour, followed a testing protocol, and guided students through ScienceSketch. Twenty-eight U.S. students in the third (N = 2), fourth (N = 17), fifth (N = 8), or sixth (N = 1) grades (aged 8–12 years) participated in the study. The study was approved by the North Carolina State University ethics board. Prior to participating, students' parents or guardians were provided informed consent and students completed an assent form. Sessions were audio and video recorded.



Fig. 1. Examples of student drawn science models for NGSS PE 4-LS1-2.

During the sessions, students completed a warm-up drawing, then learned about the fourth grade NGSS Performance Expectation (PE) 4-LS1-2: *"Use a model to describe that animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways."* Students then completed up to three modeling activities for which they could create multiple models (Fig. 1). In each activity, students saw instructions, e.g., *"Imagine you walked into a room and saw a toy. You pick it up. Draw a model to show how you sensed the toy and made the decision to pick it up."* Below the instructions was a space to draw a model using a set of drawing tools. The tools included a pen, stamps (eye, ear, nose, mouth, hand, brain), and text tools to create labels.

Evidence-Centered Design (ECD; [24]) was used to develop an initial rubric reflecting students' understanding of the NGSS PE. ECD maps learning goals and student artifacts to gauge degree of student learning. A former science teacher identified the *Competency Model* used in our ECD analysis through examining the NRC Framework [4] and the evidence statements associated with the PE, which included components and connections between them (Table 1). Three researchers then worked with the teacher to identify the *Evidence Model*, or instantiations of the *Competency Model*. The three researchers used the rubric to score an initial set of 25 models, discussed discrepancies, then updated the rubric [25]. The second set of 37 models was scored by one of two trained coders who discussed scoring decisions and finalized the scoring for all criteria.

Then, 10 models were double-scored for reliability. Across criteria, the average κ was 0.84.

Table 1. Scoring Rubric

Type	Construct	Criterion	Description
Component	Information from the environment (e.g., a flower)	(1)	Drawn information
		(2)	Typed information
Component	Sense receptor (e.g., eye)	(3)	Stamped or drawn sense receptor
		(4)	Typed sense receptor word
Component	Perception (e.g., see)	(5)	Drawn thought bubble or other indication of perception
		(6)	Typed perception word
Component	Brain	(7)	Stamped or drawn brain
		(8)	Typed word "brain"
Component	Response or action (e.g., pick up)	(9)	Stamped or drawn response
		(10)	Typed action word or description
Component	Processing occurs in the brain	(11)	Typed "process," "processing," or similar (e.g., "managing")
Connection	Information has connection to receptor	(12)	Typed connection (e.g., "detects")
		(13)	Drawn connection (line, arrow)
Connection	Receptor has a connection to the brain	(14)	Typed connection (e.g., "sends")
		(15)	Drawn connection (line, arrow)
Connection	Brain has a connection to the action	(16)	Typed connection
		(17)	Drawn connection (line, arrow)
Other	Extra information included in the model	(18)	Too many stamps or lines that obscure the model or make it challenging to understand
Other	Misconception that perception is happening in the sense receptor instead of the brain	(19)	E.g., the eye "sees" the flower instead of writing that the brain sees the flower

Per criterion, models were scored "1" if the criterion was present and "0" if the criterion was absent.

We then investigated the effectiveness of automated ratings utilizing multimodal LLMs, which are capable of processing and generating multiple types of data including text, images, audio, and video. This work investigates two such LLMs: GPT-4o from

OpenAI [26], and Gemini 2.0 Flash from Google [27]. Student images are passed to the LLMs via a web-based API, along with a natural language prompt describing the role the LLM should assume, the desired output format, as well as instructions for how to rate each of the rubric's criteria (Fig. 2). The prompt underwent several rounds of revision based on its performance assessing the initial set of 25 student science models, primarily focused on expanding the instructions and providing short examples for criteria where it struggled and incorporating practices from the OpenAI prompt engineering guide [28].

For the OpenAI model, images and prompts were passed to the LLMs using the Completions API, with images encoded as Base64 payloads. For the Google LLM, the images were sent as PNG files using the Python Imaging Library (Pillow) and the Google genai Python API. For each LLM, all other parameters utilized their default values, with the exception that the maximum response tokens was set to 1,000.

You are an elementary school science teacher evaluating a student-generated drawing. Students created science models based on the following concept: "Use a model to describe that animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways."
…..
Scoring Criteria (19 Total):

1. Drawn Stimulus: Contains a drawn freehand object from the environment that is being sensed. The stimulus must be a drawn freehand object from the environment (e.g., a flower, a sound wave, a scent particle) that the sense receptor interacts with. Do not count the sense receptor itself (e.g., eye, ear, nose, hand) as the stimulus. The stimulus must clearly represent information from the environment that the model aims to sense.
2. Typed Stimulus Description: Does the drawing contain a written description of a stimulus that is being sensed? The student typed/labeled some information from the environment. The information does NOT count if it is only in the action response section. But, if it's used in some other place, even if it is generic or vague (ex. "it" "that") then it works.
3. Drawn Sense Receptor: Includes a drawn or stamped sense receptor (e.g., eye, ear, nose, hand, tongue).
4. Typed Sense Receptor: Is there a textual representation of a sense receptor?
5. Drawn Perception: Represents perception visually, ideally in or near the brain (e.g., thought bubbles showing "seeing").
6. Typed Perception Word: Includes typed perception words like "hear," "see," or "perceive."
7. Drawn Brain: Includes a drawn or stamped brain.
8. Typed Brain: Mentions the word "brain" anywhere in the text.
…..

Fig. 2. Excerpt of *Revised Prompt* for LLMs.

4 Results

Multimodal LLM-based assessment models were evaluated based on the human scores for all 62 science models. Table 2 presents classification results for each LLM-based model in terms of accuracy and Cohen's κ for each criteria. Models varied based on both the prompt and the multimodal LLM used. *Initial Prompt* refers to the early version of the prompt that only returned ratings for each image and had short instructions for each criterion. *Revised Prompt* refers to the expanded prompt created through several iterations of prompt engineering, primarily focused on expanding the instructions for criteria where the model performed poorly. Overall, prompt revisions improved the average accuracy of the models 3% and .1 κ for the GPT-4o-based models, and 10% and .2 κfor the Gemini Flash-based models.

The largest impact on performance was between the two different multimodal LLMs investigated. The GPT 4o model performed well with the *Revised Prompt* achieving an accuracy of 77.7% and $\kappa = .46$. The Google Gemini Flash 2.0-based models performed the best, with the *Revised Prompt* variant achieving an average accuracy of 82.3% and $\kappa = .58$. This model also showed good (>.6) agreement on 10 of the criteria.

The multimodal LLMs performed well in assessing the student drawings, particularly given that they were not pre-trained with any annotated drawings from this dataset. For the first eleven criteria, which focus on components of the drawing, the best performing LLM-based model achieved a κ of >.6 for all except two criteria (one of which had κ = .59), with an average accuracy of 88%. This is likely due to these criteria being less dependent on relationships between multiple objects. Some criteria, such as Criterion 6 and Criterion 18, are difficult to evaluate due to the imbalanced nature of the ratings (almost all 1's or all 0's). The model performed notably worse on the "Connection," "Extra," and "Misconception" criteria. Connections are often defined by freehand strokes between two objects, making them difficult to interpret. In general, the models tend to be overly generous when rating Connections, with much of the errors coming from false positives. Extra information and Misconceptions were notably difficult for human raters and required discussion to refine the rubric before agreement was reached.

Limitations of the work include having a relatively small sample size of student generated drawings, and the fact that the drawings were collected outside of a classroom context (i.e., during sessions moderated by a researcher). Additionally, all the multimodal LLMs evaluated are proprietary models, and could present cost issues when attempting to scale the system to a larger number of users. The assessment framework also requires the system to make a web request to the LLM's API, which is dependent on classroom network bandwidth to allow the approach to work in a timely manner.

Table 2. Accuracy and κ by criterion for each model

Criteria																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Avg
***Initial Prompt* GPT-4o**																				
Acc	70	79	98	66	70	87	93	83	69	80	75	50	69	67	69	74	75	91	45	74
κ	.26	.44	.79	.32	.35	0	.76	.64	.37	.51	.32	.12	.4	.36	.39	.48	.44	0	0	.36
***Revised Prompt* GPT-4o**																				
Acc	87	79	100	70	61	87	96	88	66	87	95	64	58	72	69	69	72	95	56	77
κ	.73	.5	1	.45	.32	.18	.87	.77	.32	.63	.77	.16	.24	.42	.4	.37	.44	0	.09	.46
***Initial Prompt* Gemini Flash 2.0**																				
Acc	69	85	90	75	85	77	96	85	64	96	30	45	64	54	77	62	70	90	45	72
κ	.21	.62	.37	.51	.65	0	.86	.69	.23	.9	.06	.16	.32	.25	.55	.33	.43	.2	0	.38
***Revised Prompt* Gemini Flash 2.0**																				
Acc	95	87	96	67	79	98	93	85	83	95	91	48	72	70	77	87	70	95	66	82
κ	.89	.68	.65	.4	.59	.66	.78	.67	.67	.84	.72	.14	.46	.43	.55	.75	.43	.38	.33	.58

5 Conclusion

Learning-by-drawing offers significant potential for improving learning, particularly in areas such as elementary grade science. However, to fully realize the potential of drawing-based activities, students need to have proper levels of support while creating and revising

their drawings. To enable such support, we investigated the effectiveness of multimodal LLMs to understand and assess student drawn science models. First, we developed a rubric to identify how students could communicate their conceptual understanding in drawings and used it to rate the student models. Then, we developed prompts and evaluated the ability of two multimodal LLMs to assess the student drawings in relation to the human scoring. While prompt engineering produced some improvements, the largest differences in ratings were due to which LLM was used. The best performing LLM used the Google Gemini Flash 2.0 model, and was able to assess the student drawn science models with an average accuracy of 82.7% and a κ of .58 compared to the human scoring, indicating the promise of this approach.

Acknowledgements. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A210358 to North Carolina State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fiorella, L., Mayer, R.E.: Learning as a Generative Activity. Cambridge University Press (2015)
2. Leutner, D., Schmeck, A.: The generative drawing principle in multimedia learning. In: The Cambridge Handbook of Multimedia Learning. Cambridge University Press (2014)
3. Cromley, J.G., Du, Y., Dane, A.P.: Drawing-to-learn: does meta-analysis show differences between technology-based drawing and paper-and-pencil drawing? J. Sci. Educ. Technol. **29**, 216–229 (2020)
4. Council (NRC), N.R.: A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. The National Academies Press, Washington, DC (2012)
5. States, N.L.: Next Generation Science Standards: For States, By States. The National Academies Press, Washington, DC (2013)
6. Schmeck, A., Mayer, R.E., Opfermann, M., Pfeiffer, V., Leutner, D.: Drawing pictures during learning from scientific text: testing the generative drawing effect and the prognostic drawing effect. Contemp. Educ. Psychol. **39**, 275–286 (2014)
7. Schwamborn, A., Thillmann, H., Opfermann, M., Leutner, D.: Cognitive load and instructionally supported learning with provided and learner-generated visualizations. Comput. Hum. Behav. **27**, 89–93 (2011)
8. Wiebe, E.N., Bedward, J.C., Madden, L.P.: Graphic representations in science notebooks: a vehicle for understanding science inquiry in the elementary classroom. In: Proceedings of AERA. AERA (2009)
9. Forbus, K.D., Chang, M., McLure, M., Usher, M.: The cognitive science of sketch worksheets. Top. Cogn. Sci. **9**, 921–942 (2017)
10. Miller, B., Cromley, J., Newcombe, N., Chang, M.D., Forbus, K.D.: Using CogSketch to support student science learning through sketching with automatic feedback. In: Proceedings of AERA. AERA (2014)

11. Yin, P., Forbus, K., Usher, J., Sageman, B., Jee, B.: Sketch worksheets: a sketch-based educational software system. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1871–1876 (2010)
12. Van Joolingen, W.R., Schouten, J., Leenaars, F.: Drawing-based modeling in teaching elementary biology as a diagnostic tool. In: Upmeier zu Belzen, A., Krüger, D., van Driel, J. (eds.) Towards a Competence-Based View on Models and Modeling in Science Education, pp. 131–145. Springer (2019)
13. Shute, V., Almond, R., Rahimi, S.: Physics Playground (1.3) [Computer software] (2019). https://pluto.coe.fsu.edu/ppteam/pp-links/
14. Shute, V.J., Rahimi, S., Smith, G.: Game-based learning analytics in physics playground. In: Chang, M., Tlili, A. (eds.) Data Analytics Approaches in Educational Games and Gamification Systems, pp. 69–93. Springer, New York (2019)
15. Shute, V., et al.: The design, development, and testing of learning supports for the Physics Playground game. Int. J. Artif. Intell. Educ. **31**, 357–379 (2020)
16. Viswanathan, V.K., Hurt, J.T., Hammond, T., Caldwell, B., Talley, K., Linsey, J.: Impact of a sketch-based tutoring system at multiple universities. In: ASEE Virtual Annual Conference Content Access (2020)
17. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. J. Res. Sci. Teach. **59**, 1765–1794 (2022)
18. Li, T., Liu, F., Krajcik, J.: Automatically assess elementary students' hand-drawn scientific models using deep learning of artificial intelligence. In: ISLS Annual Meeting 2023, pp. 1813–1814 (2023)
19. Henkel, O., Hills, L., Roberts, B., McGrane, J.: Can LLMs grade open response reading comprehension questions? An empirical study using the ROARs dataset. Int. J. Artif. Intell. Educ. 1–26 (2024)
20. Cohn, C., Hutchins, N., Le, T., Biswas, G.: A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 23182–23190 (2024)
21. Kortemeyer, G.: Performance of the pre-trained large language model GPT-4 on automated short answer grading. Discov. Artif. Intell. **4**, 47 (2024)
22. Kim, Y.J., et al.: Dual process masking for dialogue act recognition. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 15270–15283 (2024)
23. Acosta, H., et al.: Recognizing multi-party epistemic dialogue acts during collaborative game-based learning using large language models. Int. J. Artif. Intell. Educ. 1–25 (2024)
24. Mislevy, R.J., Haertel, G.D.: Implications of evidence-centered design for educational testing. Educ. Meas. Issues Pract. **25**, 6–20 (2006)
25. McDonald, N., Schoenebeck, S., Forte, A.: Reliability and inter-rater reliability in qualitative research: norms and guidelines for CSCW and HCI practice. Proc. ACM Hum.-Comput. Interact. **3**, 1–23 (2019)
26. OpenAI: GPT-4o [Large language model] (2024)
27. DeepMind, G.: Gemini Flash 2.0. Google (2025)
28. OpenAI.: Best Practices: Prompt Engineering (2024)

Riding on the Back of a Whale: A Hackathon Framework for Introducing High School Students to Large Language Models

Duc Nguyen[1], Dong Le[1], Long Nguyen[1], Quyen Vu[1], Tran Le[1], Dung Nguyen[1], Nga Huynh[1], Huong Nguyen[1], Phat Tran[1], Dang Le[1], Sang Truong[2], Sanmi Koyejo[2], Cuong Le[3], and Tho Quan[1(✉)]

[1] Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam
{nqduc,dong.le0110,long.nguyencse2023,quyen.vu12, tran.lethanhbao,dung.nguyennhdrn,nga.huynhnga,huong.nguyenphuocngoc, phatttt,binhdang,qttho}@hcmut.edu.vn
[2] Stanford University, Stanford, CA 94305, USA
{sttruong,sanmi}@cs.stanford.edu
[3] Ton Duc Thang University, Ho Chi Minh City, Vietnam
leanhcuong@tdtu.edu.vn

Abstract. As large language models (LLMs) become more integrated into daily life, it is crucial to foster AI literacy among high school students. However, most AI courses target college-level learners and assume prior knowledge, while high schools often lack the foundational curriculum and infrastructure for traditional LLM education. To bridge this gap, we present a hackathon-based framework that makes LLM learning accessible, engaging, and hands-on. The program combines interactive lectures on core LLM concepts with a guided competition where students fine-tune models and build real-world applications, such as healthcare chatbots. This approach boosts motivation, programming skills, and practical understanding. Post-hackathon survey results show students gained both functional LLM experience and foundational knowledge. Furthermore, our framework can be extended to broader audiences, including learners without prior AI/NLP experience, offering a rapid, application-driven introduction to LLMs.

Keywords: AI Literacy · Hackathon-Based Learning · Large Language Models Education

1 Introduction

Education has long been a cornerstone of society, with AI education becoming increasingly vital in recent years. AI literacy is recognized as a crucial skill

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 201–209, 2025.
https://doi.org/10.1007/978-3-031-99264-3_25

for students in the 21^{st} century [14,18,19]. As AI technologies rise in prominence, integrating AI education into K-12 schools—particularly high schools—has gained momentum to prepare students for an AI-driven future [4,10,19,23]. Students should not only understand AI concepts but also gain practical skills for solving problems and designing solutions [17,18,25]. With the rise of LLMs, we aim to provide high school students aspiring to enter tech fields with foundational knowledge and skills in LLMs. Traditionally, LLM curricula are designed for university-level students and require strong math and computer science backgrounds, making them less accessible to younger learners. While AI education is being integrated into K-12 curricula, studies on LLM-specific education are limited [20,27,28]. Furthermore, these curricula often lack customization for high school students. Modern teaching methods like collaborative, project-based, and game-based learning enhance engagement and social skills [17]. Recent hackathons and competitions combine these methods but often provide only basic exposure to generative AI, especially LLMs. To improve LLM education for high school students, we propose a structured hackathon with interactive lectures and guided competition, reinforcing learning through hands-on application.

Our contributions include: (i) We develop comprehensive educational resources, including lecture slides and instructional videos, to teach foundational LLM concepts, fine-tuning techniques, and required knowledge to deal with LLM limitations; (ii) We design detailed, step-by-step, hands-on tutorials with accompanying code, enabling students to build a functional LLM application; and (iii) We propose a structured hackathon model specifically designed to engage high school students in LLM applications, fostering both AI literacy and hands-on problem-solving skills.

2 Related Works

Pedagogical Approaches for Teaching Artificial Intelligence in High School. Effective AI education often uses practical approaches such as collaborative, project-based, and game-based learning [17]. Collaborative learning builds teamwork and critical thinking through group problem-solving [3,13,14]. Project-based learning emphasizes solving real-world problems via structured projects [7–9,16]. Game-based learning boosts motivation and understanding through gameplay and gamification [6,12,15,22,26]. Recent AI education efforts have merged pedagogical strategies into hackathons, where students build AI solutions in teams and compete for prizes. For instance, [1] used Scratch to teach explainable AI, while [21] offered a month-long ML and NLP course with projects. However, LLM-focused instruction in high schools remains limited due to infrastructure challenges and the difficulty of simplifying complex concepts, requiring both pedagogical and technical expertise.

Large Language Model Competitions for High School Students. AI competitions enhance high school education by linking theory to real-world practice [2]. Online

ones like Microsoft's Imagine Cup Junior and ColorShapeLinks [5] increase engagement. While many programs promote AI literacy, few focus on LLMs. WAICY includes an LLM track but centers on prompt engineering without structured learning or mentorship. In contrast, our competition provides guided mentorship, in-depth courses on prompting, optimization, and AI ethics, plus curated materials to ensure both theoretical and practical understanding of LLMs.

3 Large Language Model Hackathon Framework

3.1 Learning Outcomes

Our competition serves as an educational means, providing participants with specific learning outcomes. By the end of the competition, students will achieve the following *Learning Outcomes* (LO): (*LO1*) Understand fundamental LLM concepts; (*LO2*) Be able to implement an LLM-powered chatbot; (*LO3*) Well collaborate in a diverse inter-school team; and (*LO4*) Develop awareness of practical LLM applications. Notably, *LO4* is intended to inspire students to pursue further learning in LLMs and AI, fostering their interest in higher education.

3.2 Material Description

Our program introduces LLM fundamentals through a structured curriculum of lectures, activities, and a hackathon. We developed necessary materials for teaching LLMs tailored to high school students, including lecture slides, instructional videos, and a hands-on guide for the hackathon, with sample code and interactive notebooks. These resources enable students to immediately experiment with LLMs and increase engagement. All these materials are available at OSF Link. Topics were selected based on the intended learning outcomes, covering LLM basics in *L01*, Retrieval-Augmented Generation (RAG) techniques [11] in *L02*, and LLM training and fine-tuning for domain-specific applications in *L04*. Preparation of the materials began six weeks prior to the hackathon. One week before the event, we conducted rehearsals among team members, including first- to fourth-year undergraduates and Computer Science (CS) professors. Based on these rehearsals, we adjusted the content to fit within the time constraints and removed advanced concepts that required a strong CS background and did not align with the learning objectives.

3.3 Preparation for the Hackathon

Mentor Team. We recruit senior undergraduates as mentors based on team numbers. Each team gets a dedicated mentor selected for strong academics and LLM experience. Mentors must have an *A* in a Machine Learning course and at least three months of LLM experience via projects or thesis work. Selection is based on a composite score (Eq. 1).

$$score = 0.6 \times course_score + 0.4 \times \left(\frac{5}{3}\right) \times months \tag{1}$$

In (1), the *course_score* is scaled to a 10-point scale with the following standards: $A^+ = 10$, $A = 9$, $A^- = 8$, $B^+ = 7.5$, $B = 7$, $B^- = 6.5$; and *months* represents the number of months the mentors have experience with LLMs.

Invited Speakers. We invited four distinguished experts, including academic scholars and industry practitioners, to deliver presentations on LLM topics. These presentations are structured into four lectures, addressing the following key areas: an overview of LLMs, RAG, recent advancements and emerging techniques in LLM research, and a hands-on tutorial focused on the fine-tuning and deployment of LLMs.

Awards and Prizes. Awards have been established to recognize the top-performing teams in the competition. These awards comprise four distinct categories: First Prize, Second Prize, Third Prize, and two Honorable Mentions. The monetary value or nature of these awards is subject to the allocated organizing budget and may vary based on the geographical location or the scale of the hackathon.

Access to LLM APIs. For the competition, participants are granted access to two LLM APIs to facilitate experimentation, dataset creation, and comparative analysis against their own LLM implementations. The first API provided is GPT-4o, hosted on the Azure platform, with each team allocated a budget of $50 for API usage. The second API is our in-house developed LLM, which is made available to students without usage restrictions. To ensure efficient API management, LiteLLM is employed as a proxy system to handle API calls.

3.4 Executive Plan

We propose a two-day hackathon with a structured learning experience through lectures, teamwork, and hands-on application. On the first day, activities begin with two lectures on introducing LLM fundamentals and RAG. After the lunch break, the students come to interactive lectures, in which they gains practical knowledges about how LLM works and how to finetune an LLM. The day concludes with the official launch of the hackathon, during which students form teams of 4âĂŞ6 to tackle real-world challenges. The second day starts with a project pitching session where student teams present their ideas, receive feedback, and refine their solutions. In the afternoon, final presentations are evaluated by a panel based on technical execution, creativity, and impact (scoring rubric). Awards are given for outstanding projects, and the event concludes with a gala dinner.

4 Results

4.1 Learning Material Evaluation

In this section, we compare the effectiveness of our teaching materials with other well-known AI educational resources. Since the criteria for evaluating AI teaching

Table 1. Evaluation of Teaching Materials

Course	Easy Adaptation	Localization	Examples & Exercises	Clarity & Coverage	Age-Suitability
Generative AI with Large Language Models	3	5	4	4	2
Generative AI Engineering with LLMs Specialization	3	5	5	5	2
Our materials	5	5	5	4	4

Table 2. Comparison of our competition with the others

Competition	Content	Time	Mentors	Place
ISEF	STEM	Over months	Professionals	Hybrid
IOAI	AI	Several days	AI Experts	In-person
WAICY	AI	Several days	–	Hybrid
BLAST AI	AI	8 weeks	Volunteers	Hybrid
HackNYU	Beyond STEM	48 h	Professionals	In-person
Hack the North	STEM	36 h	Volunteers	In-person
MHacks	Beyond STEM	36 h	Volunteers	In-person
NASA Space Apps	STEM	48 h	Professionals	In-person
PennApps	Engineering	48 h	Professionals & Graduate students	In-person
Ours	LLMs only	36 h	Well-selected undergraduates	Hybrid

materials are not extensively studied, we draw on the framework proposed by [24] to establish our own set of evaluation criteria.

We selected two online courses on Coursera for comparison, including *Generative AI with Large Language Models*, and *Generative AI Engineering with LLMs Specialization*. The selected courses originate from well-known organizations (IBM, AWS, and DeepLearning.AI), have more than 4.5 stars, target people without a strong engineering background, and have been studied more than 5,000 times. Our evaluation of these courses is presented in Table 1. In summary, our teaching materials are designed to be clear and accessible for high school students, incorporating a variety of examples and exercises to enhance understanding. However, a slight limitation lies in the coverage of LLM topics, as the materials were designed for hackathon-based teaching. This approach prioritizes quick knowledge acquisition without overwhelming students, resulting in a focused and less comprehensive coverage.

4.2 Competition Analysis

We compare our competition with three leading high school competitions, ISEF, IOAI, and WAICY, and six well-established high school hackathons: BLAST AI, HackNYU, Hack the North, MHacks, NASA Space Apps, and PennApps. The

comparison focuses on key criteria related to content and organizational structure. Table 2 highlights the distinguishing features of our hackathon. Unlike ISEF, IOAI, and NASA Space Apps, which cover a broad range of STEM and AI topics, our hackathon is uniquely dedicated to LLMs, offering a specialized and in-depth exploration of this emerging field. Structurally, it aligns with established competitions and hackathons by adopting a 36-hour format, hybrid participation model, and mentorship framework. Similar to HackNYU and PennApps, our hackathon provides expert guidance, though with a targeted focus on LLMs. This specialization, combined with a hybrid structure and structured mentorship, establishes our competition as a novel and impactful platform for engaging with LLM-related challenges.

4.3 Implementation Results

We implemented the hackathon as 2024 CSE Summer School at University of Technology - VNU-HCM, selecting 50 students from 108 registrants, while ensuring diversity in gender, grade, and schools. Six months after the hackathon, we conducted a survey and received 36 responses from the participating students, yielding a response rate of 72%. The anonymized student list, submissions, survey questions, and results are available at OSF Link. We summarize the survey results in Table 3. The cohort consisted of 76% male, 24% female, with 72% in Grade 12, 24% in Grade 11, and 4% in Grade 10, from 29 schools. Students were assigned to 10 groups, each with a mentor.The follow-up survey after six months showed a gap between theoretical knowledge (50% intermediate understanding) and confidence (25% able to explain LLM principles). Fine-tuning and prompt engineering were well understood (72.2%), while bias and ethics were less familiar (22.2%). Most students (83.4%) implemented a chatbot but faced challenges with response accuracy (75%), with 83.3% believing prompt modifications could help. Time management (50%) and communication barriers (27.8%) were the biggest difficulties. Despite this, 77.7% had positive teamwork experiences. After the hackathon, 72.2% reported high awareness of LLM applications, and 88.8% expressed interest in further LLM study, highlighting the event's success.

Table 3. Survery results

Questions	Summarized Results
Q1: What is your current level of familiarity with Large Language Models (LLMs)?	5.6% no knowledge; 41.7% basic knowledge; 50.0% intermediate knowledge; 2.8% advanced knowledge
Q2: How confident are you in explaining the basic working principles of LLMs?	5.6% not confident; 25% somewhat confident; 44.4% neutral; 19.4% confident; 5.6% very confident
Q3: Which key concepts related to LLMs do you feel most comfortable with? (multiple answers)	55.5% tokenization; 66.6% training and inference; 72.2% fine-tuning and prompt engineering; 22.2% bias and ethical considerations

continued

Table 3. continued

Questions	Summarized Results
Q4: How well do you understand the strengths and limitations of LLMs?	5.6% poorly; 22.2% somewhat well; 63.9% moderate well; 8.3% very well
Q5: Have you successfully implemented a basic chatbot using an LLM?	5.6% Yes, independently; 77.8% Yes, with helps from peers/mentors; 8.3% No, but I understand the process; 8.3% No, and I need further clarification
Q6: Which aspects of chatbot implementation did you find most challenging? (multiple answers)	41.7% handling user inputs effectively; 75% improving response accuracy; 41.7% managing biases and ethical concerns; 41.7% understanding API integration
Q7: How comfortable are you with modifying prompts to improve chatbot responses?	5.6% not comfortable; 2.8% somewhat comfortable; 8.3% neutral; 69.4% comfortable; 13.9% very comfortable
Q8: How would you rate your experience working in a team on this project?	0% very negative; 2.8% somewhat negative; 19.4% neutral; 44.4% somewhat positive; 33.3% very positive
Q9: What challenges did you encounter while collaborating with your team? (multiple answers)	27.8% communication issues; 27.8% technical disagreements; 19.4% task delegation difficulties; 50.0% time management problems; 2.8% other
Q10: How well did your team manage to distribute workload and responsibilities?	0% Poorly; 27.8% somewhat well; 41.7% moderately well; 30.6% very well
Q11: After completing this project, how would you rate your awareness of real-world applications of LLMs?	2.8% very low; 5.6% somewhat low; 19.4% neutral; 38.9% somewhat high; 33.3% very high
Q12: How interested are you in further exploring or applying LLMs in your studies/career?	0% not interested; 0% slightly interested; 11.1% neutral; 44.4% interested; 44.4% very interested

5 Conclusion

In this work, we present a hackathon framework designed to introduce high school students to the fundamentals of LLMs and provide hands-on experience in developing and applying LLMs to real-world problems. Through lectures and a competition, students gain essential knowledge of LLM design, techniques, and applications, fostering creative problem-solving and bridging theoretical understanding with practical implementation. The program encourages active engagement, deepens comprehension, and helps students explore future career directions in the field. We believe this hackathon equips learners with the fundamental skills and knowledge necessary to effectively use and manage LLMs in their future studies and careers, akin to the ability to "*ride the whale*" of this transformative technology.

Acknowledgements. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study. SK acknowledges support by NSF 2046795, NSF 2205329, and IES R305C240046. The GPT-4o API is supported by the Microsoft Accelerating Foundation Models Research program. We also acknowledge DSciLab (HCMUT-VNU-HCM) for providing partial computing resources used in this study.

References

1. Alonso, J.M.: Teaching explainable artificial intelligence to high school students. Int. J. Comput. Intell. Syst. **13**, 974–987 (2020)
2. Carpio Cañada, J., Mateo Sanguino, T., Merelo Guervós, J., Rivas Santos, V.: Open classroom: enhancing student achievement on artificial intelligence through an international online competition. J. Comp. Assist. Learn. **31**(1), 14–31 (2015)
3. Carpio Cañada, J., Mateo Sanguino, T., Merelo Guervós, J., Rivas Santos, V.: Open classroom: enhancing student achievement on artificial intelligence through an international online competition. J. Comput. Assist. Learn. **31**(1), 14–31 (2015)
4. Chai, C.S., Lin, P.Y., Jong, M., Dai, Y., Chiu, T.K., Qin, J.: Perceptions of and behavioral intentions towards learning artificial intelligence in primary school students. Educ. Technol. Soc. **24**(3), 89–101 (2021)
5. Fachada, N.: ColorShapeLinks: a board game AI competition for educators and students. Comput. Educ. Artif. Intell. **2**, 100014 (2021)
6. Kapp, K.: The gamification of learning and instruction: game-based methods and strategies for training and education. Pfeiffer (2012)
7. Kong, S.C., Man-Yin Cheung, W., Zhang, G.: Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. Comput. Educ. Artif. Intell. **2**, 100026 (2021)
8. Krajcik, J.S., Blumenfeld, P.C.: Project-based learning. In: Cambridge Handbooks in Psychology, pp. 317–334. Cambridge University Press, Cambridge (2005)
9. Kumar, A., Kumar, D., Russell, I.: Non-traditional projects in the undergraduate ai course. In: Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '06, pp. 479–480. Association for Computing Machinery, New York (2006)
10. Lee, S.J., Kwon, K.: A systematic review of AI education in K-12 classrooms from 2018 to 2023: topics, strategies, and learning outcomes. Comput. Educ. Artif. Intell. **6**, 100211 (2024)
11. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook (2020)
12. Li, Z., O'Brien, L., Flint, S., Sankaranarayana, R.: Object-oriented sokoban solver: a serious game project for ooad and ai education. In: 2014 IEEE Frontiers in Education Conference (FIE) Proceedings, pp. 1–4 (2014)
13. van der Linden, J., Erkens, G., Schmidt, H., Renshaw, P.: Collaborative Learning, pp. 37–54. Springer, Dordrecht (2000)
14. Long, D., Magerko, B.: What is ai literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, pp. 1–16. Association for Computing Machinery, New York (2020)
15. McGovern, A., Tidwel, Z., Rushing, D.: Teaching introductory artificial intelligence through java-based games. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 25, no. 3, pp. 1729–1736 (2021)
16. Michaud, L.N.: Evil robots and helpful droids: a seminar for a junior/senior artificial intelligence course. J. Comput. Sci. Coll. **29**(6), 123–129 (2014)
17. Ng, D., Lee, M., Tan, R., Hu, X., Downie, J.S., Chu, S.: A review of ai teaching and learning from 2000 to 2020. Educ. Inf. Technol. **28**(7), 8445–8501 (2023)
18. Ng, D., Leung, J., Chu, K., Qiao, M.S.: AI literacy: definition, teaching, evaluation and ethical issues. Proc. Assoc. Inf. Sci. Technol. **58**(1), 504–509 (2021)

19. Ng, D.T.K., Leung, J.K.L., Su, M.J., Yim, I.H.Y., Qiao, M.S., Chu, S.K.W.: AI Literacy in K-16 Classrooms. Springer, Heidelberg (2022)
20. Ng, D., Luo, W., Chan, H., Chu, S.: Using digital story writing as a pedagogy to develop ai literacy among primary students. Comput. Educ. Artif. Intell. **3**, 100054 (2022)
21. Norouzi, N., Chaturvedi, S., Rutledge, M.: Lessons learned from teaching machine learning and natural language processing to high school students. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 09, pp. 13397–13403 (2020)
22. Ribeiro, P., Simões, H., Ferreira, M.: Teaching artificial intelligence and logic programming in a competitive environment. Inf. Educ. **8**(1), 85–100 (2009)
23. Rizvi, S., Waite, J., Sentance, S.: Artificial Intelligence teaching and learning in K-12 from 2019 to 2022: a systematic literature review. Comput. Educ. Artif. Intell. **4**, 100145 (2023)
24. Tomlinson, B.: Developing principled frameworks for materials development. In: Developing Materials for Language Teaching, pp. 107–129 (2003)
25. Touretzky, D., Gardner-McCune, C., Seehorn, D.: Machine learning and the five big ideas in AI. Int. J. Artif. Intell. Educ. **33**(2), 233–266 (2023)
26. Wong, D., Zink, R., Koenig, S.: Teaching artificial intelligence and robotics via games. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 24, no. 3, pp. 1917–1918 (2010)
27. Xia, Q., Chiu, T.K., Lee, M., Sanusi, I.T., Dai, Y., Chai, C.S.: A self-determination theory (sdt) design approach for inclusive and diverse artificial intelligence (ai) education. Comput. Educ. **189**, 104582 (2022)
28. Yang, W.: Artificial intelligence education for young children: why, what, and how in curriculum design and implementation. Comput. Educ. Artif. Intell. **3**, 100061 (2022)

Design and Validation of the Psychometric Properties of a Questionnaire Measuring AI Literacy Among University Students: Preliminary Results

Chantal Tremblay[1](✉), Bruno Poellhuber[2], Anne-Marie Sénécal[2], and Normand Roy[2]

[1] Université du Québec à Montréal, Montréal, QC H3C 3P8, Canada
tremblay.chantal@uqam.ca
[2] Université de Montréal, Montréal, QC H3T 1J4, Canada

Abstract. The rapid adoption of digital technologies and artificial intelligence (AI) necessitates equipping university students with AI literacy. Since the rise of Generative AI in 2020, frameworks have emerged to define key competencies, spanning technical knowledge, ethical considerations, and AI applications. This paper presents an overview of an ongoing research project assessing undergraduate students' perceived self-efficacy in AI literacy in Quebec. Specifically, it focuses on the design and validation of the questionnaire's psychometric properties. The study follows a rigorous methodological approach, grounded in a scoping review of AI literacy frameworks, which identified five key dimensions: Technical aspects, Ethical aspects, Critical thinking, AI use and Human agency.

Study 1 involved 768 students completing an initial questionnaire in 2023. It encompassed three key dimensions—Technical aspects, Ethical considerations, AI use and general computer literacy. Exploratory factor analysis (EFA) with Varimax rotation validated a four-factor structure, explaining 57.5% of variance across 28 items. **Study 2** expanded the questionnaire to integrate dimensions not covered in Study 1 (AI use for learning, Critical thinking, Human agency). Twelve items from three scales in Study 1 were retained, while 34 additional items were added to ensure comprehensive coverage of all dimensions identified in the scoping review. The revised questionnaire is being administered at two universities in Quebec and will be subject to exploratory and confirmatory factor analyses. The findings will be presented at the upcoming conference in July.

Keywords: AI literacy frameworks · Higher Education · Questionnaire Development · Self-efficacy

1 Introduction: The Need for an Instrument to Measure University Students' AI Literacy

The growing adoption of digital technologies and AI in organizations reflects recent technological advancements. Companies are increasingly implementing remote work, GenAI, cloud computing, and Big Data to optimize value chains and profitability [1,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 210–217, 2025.
https://doi.org/10.1007/978-3-031-99264-3_26

2]. Digital technology is also reshaping collaboration, communication, and information retrieval [4]. Given this shift, many emphasize the need to equip university students with digital skills and AI literacy.

From 2010 to 2020, numerous digital competence frameworks emerged [3–5], including the European Commission's DigComp framework [6, 7]. Since the rise of GenAI in 2022 AI literacy frameworks have followed, aiming to define the skills students need, from technological knowledge to AI usage for various purposes [8–11]. Many also address broader societal implications, particularly those linked to AI-driven transformations. UNESCO's AI literacy framework [11] highlights concerns, such as representational biases affecting equity, diversity, and inclusion, as well as issues related to human agency and AI's environmental impact.

In 2019, Quebec's Ministry of Education introduced the Digital Competency Framework [12], developed from a review of over 70 international frameworks on digital competence, information literacy, and 21st-century skills [3]. The framework outlines twelve interrelated dimensions, covering ethical digital behavior, technological skills, collaboration, content creation, and problem-solving with digital tools. Five years later, the Ministry of Education collaborated with our research team to update this framework by incorporating AI literacy. Similar to many education systems [11], Quebec currently lacks an official AI literacy framework.

In parallel with this work, we initiated a project to measure the self-efficacy beliefs of university students in Quebec regarding their AI literacy. This project was carried out in two phases, the first of which took place in 2023 and focused on designing and validating the psychometric properties of a questionnaire intended to assess certain dimensions of AI literacy (Study 1). Building on the scoping review of AI literacy frameworks conducted to update the Quebec framework, the second and current phase of the project aims to design and validate the psychometric properties of the revised version of this questionnaire, which now covers all dimensions of AI literacy (Study 2).

The need for this instrument and study arises from the lack of recent data, as the last large-scale assessments of digital skills in Quebec date back to 2011[13]. Saltos-Rivas et al. [14] emphasize the importance of using validated instruments tailored to the target population, a principle that also applies to AI literacy. While existing AI literacy questionnaires (e.g., [15, 16]) are available, we found it important to develop an instrument that aligns with the Quebec framework. Thus, all dimensions of AI literacy will be integrated, in the same way that the Quebec Digital Competency Framework represents this competency by considering all related issues and uses. This study also seeks to address the gap in validated instruments for measuring digital competence and AI literacy [14–18]. Finally, this study seeks to address limitations found in existing questionnaires by developing an instrument that captures all dimensions of AI literacy, rather than focusing only on a subset [15].

2 Conceptual Framework: AI Literacy and Self-Efficacy

Over the past five years, AI literacy has been the focus of numerous studies aimed at defining the concept and identifying its components [8–11, 19–21]. While there is no consensus on the definition of AI literacy [8, 11], the definition proposed by Long and

Magerko [10] is widely cited in the literature. According to them, AI literacy is defined as "a set of competencies that enable individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (p. 2). They justify their definition by relying on their interpretation of the term "literacy", which they define as "skill sets in a variety of disciplines that have the potential to enable expression, communication, and access to knowledge" (p. 2). However, others have argued that there is a conceptual difference between "AI literacy" and "AI competence" [9]. According to the authors, the former refers to knowledge of AI and skills related to its use, whereas the latter encompasses AI literacy along with attitudes, beliefs, and mindsets related to the impacts of AI on society [9] In the UNESCO student framework [11], Miao and Shiohira also noted that these definitions are often influenced by private companies that provide training or access to AI technologies, which leads to an emphasis on technical skills at the expense of critical thinking and a human-centered mindset. This aligns with the findings of Almatrafi et al. [8], who show that AI literacy definitions vary depending on "the purpose, target population, domain, and field of study" (p. 5).

2.1 AI Literacy

Considering the variation in definitions of AI literacy and its components, our research team conducted a scoping review to examine the current state of knowledge on AI literacy in education. This review aimed to synthesize the key components of AI literacy frameworks and outline the knowledge, abilities, skills, or attitudes associated with each. To carry out this review, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) procedure [31]. The documents included in the review met the following inclusion criteria: they examined the concept of AI literacy in education; proposed specific AI literacy competencies or skills; involved populations closely connected to education; were published between January 2018 and April 2024; were disseminated as journal articles (theoretical or empirical works, systematic reviews) or documents (websites of institutions, associations or governments); were disseminated in English or French, except for documents published by government bodies, which were accepted in any language and translated when necessary. Out of 353 documents initially identified, title and abstract screening reduced the corpus to 90. A full-text review then further refined the selection based on inclusion criteria and relevance. After this, a final corpus of 31 documents was included in the review.

We used the conceptual framework derived from this review to refine our initial AI literacy questionnaire. Five main dimensions emerged from our analysis: Technical aspects, Ethical aspects, Critical thinking, AI use, and Human agency. The definitions of each dimension are presented in Table 1. In the coding scheme, each dimension is subdivided into themes and further into codes [supplementary material].

Although a mechanism for students to complete multiple complex tasks to assess their AI literacy and digital competence would be valuable, developing such tasks and ensuring their reliability and validity would require significant resources [14]. It would also increase the time needed for participants to complete the study. To address this feasibility challenge, an effective approach is to measure AI literacy using a questionnaire that gathers students' perceived self-efficacy (PSE) for each of these dimensions.

Table 1. Dimensions and definitions of AI literacy

Dimensions	Definitions
Technical aspects	Technical knowledge and skills related to AI and its systems
Ethical aspects	Knowledge and skills concerning ethical issues in the use and design of AI systems
Critical thinking	Knowledge and skills required to critically evaluate AI usage
AI use	Knowledge and skills necessary for the effective and appropriate use of AI systems, including learning, teaching, and personal contexts
Human Agency	Knowledge, skills, and attitudes essential for fostering and preserving human agency in human-AI interactions and AI system usage

2.2 Perceived Self-efficacy

Bandura's self-efficacy theory [22] explains the factors that motivate individuals to engage in task completion. Self-efficacy reflects a person's belief in their ability to take action and achieve desired outcomes [22]. According to this theory, the stronger an individual's belief in their capabilities, the more determined they will be to engage in tasks. In other words, the more a person perceives themselves as competent—believing they can achieve the desired results—the more committed they will be to performing the tasks necessary to attain those results. Previous studies have shown that digital self-efficacy is a key determinant of digital technology use, although actual usage does not systematically reflect the level of competence [24]. Some studies have demonstrated a relationship between perceived self-efficacy, and actual digital competence [23, 24]. Accordingly, AI literacy can be assessed by evaluating perceived self-efficacy across the five dimensions of our framework. However, it is important to keep in mind that students with low competence may overestimate their self-efficacy, a phenomenon known as the Dunning-Kruger effect [25].

3 Study 1: Initial Version of the Questionnaire

As part of a broader survey on the use of ChatGPT in a university context, the initial version of the AI literacy questionnaire collected 768 responses from undergraduate and graduate university students across all faculties. It comprised 28 items encompassing three dimensions: Technical aspects (AI technical knowledge, general computer and AI proficiency), Ethical aspects, and AI use (personal contexts).

To evaluate its psychometric properties, an exploratory factor analysis (EFA) with Varimax rotation was conducted [26, 27]. Using an eigenvalue greater than 1 and without forcing the number of factors, the analysis resulted in a four-factor solution explaining 57.5% of the variance across 27 final items. McDonald's omega coefficient was used to evaluate the internal consistency reliability [28, 29]. The twelve items related to a subscale of the first dimension, AI technical knowledge, loaded onto the same factor, with loadings between .659 and .854, with an Omega of .947. Of the five items related to General computer and AI proficiency, only three loaded onto the correct factor, with

a relatively low omega of .586, so we decided to drop it. The five items related to the second dimension, Ethical aspects, also loaded onto the same factor, with an Omega of .735. The six items related to the third dimension, AI use, loaded onto the same factor, but to ensure coherence and comparability with an equivalent questionnaire designed for teachers, only four items were retained, with an Omega of .821. Thus, our final solution in study 1 consists of on 21 items across three subscales.

4 Study 2: Refinement of the Questionnaire

Based on the findings from the AI literacy scoping review, conducted in parallel with the current project, it was determined that the initial questionnaire did not comprehensively capture all components of AI literacy. To address this limitation, the questionnaire was refined using the coding scheme from the scoping review to ensure its completeness. The revision process began by identifying the dimensions, themes, and codes that were not adequately covered in the initial version. Subsequently, 33 new items were developed to fill these gaps. First, four items were added to cover the Technical aspects dimension, specifically addressing data, AI architecture, and machine learning. Six additional items were introduced to address the Ethical aspects dimension, such as societal issues, issues related to AI system design (ethics by design), and AI regulations. Notably, two items that were initially categorized under Ethical aspects in Study 1 were reassigned to the Critical thinking dimension in Study 2, as they pertained to assessing potential risks associated with AI use in society. Furthermore, four new items were designed to encompass other themes within this dimension, including self-assessment and self-regulation of AI use, as well as critical thinking regarding AI applications. To better capture AI usage, 15 items were incorporated, focusing on learning-related AI use (n = 6) and everyday AI use (n = 9). These items cover aspects such as human-AI collaboration, AI prompting, communication with AI agents and AI-assisted problem-solving. Finally, four additional items were created to assess the Human agency dimension.

Since the questionnaire aims to measure perceived self-efficacy, each item in the revised version was formulated to begin with the statement: "*I feel competent that I can…*". Students are then asked to rate their response on a seven-point Likert scale, ranging from "*strongly disagree*" to "*strongly agree*".

These items underwent a qualitative pre-validation with five members of a research lab specializing in higher education pedagogy, with a focus on digital technology and AI. The goal was to ensure that the items accurately represent the target AI literacy dimensions and themes. Then, a group of students was consulted to provide feedback on item wording, ensuring clarity and relevance. Their input helped refine the phrasing of certain items without altering their number. Table 2 shows an example item for each dimension (Refer to the supplementary material of the paper for the full list of items.)

At the time of submitting this paper, the second phase of data collection has just ended. In this phase, 20% of undergraduate students enrolled in two major French universities in Quebec were randomly selected and received an invitation email from the institutional body responsible for student surveys. A total of 975 students (including 709 women) participated in the survey, of whom 53.4% were in their first (28.5%) or second year of university studies. A total of 62.7% of participants are between the ages of 20

Table 2. Example items for each dimension

Dimensions	Item *I feel competent that I can…*
Technical aspects	Explain to someone else how natural language processing works
Ethical aspects	Name unethical uses of AI
Critical thinking	Anticipate future impacts of AI on the environment
AI use	Use generative AI tools to solve various problems
Human Agency	Maintain control over my actions when interacting with AI

and 25. Data analysis will be conducted prior to the conference in July. First, our sample will be divided into two subsamples. For the first subsample, EFA will be run on the new AI questionnaire following the same procedure described in Study 1, running also internal consistency reliability analysis. For the second half of the sample, confirmatory factor analysis relying on structural equations in HLM. To ensure conceptual validity, the results of these analysis will be compared with the results of the scoping review and AI literacy frameworks.

5 Discussion and Conclusion

This study led to the development of a questionnaire designed to assess university students' AI literacy. It distinguishes itself from other questionnaires by conceptualizing AI literacy across five dimensions, which were developed based on the findings of a scoping review. This instrument thus enables a comprehensive and detailed assessment of self-efficacy in AI literacy. The validation of its psychometric properties on a large sample will evaluate internal consistency and aims to provide empirical evidence supporting the relevance of the dimensions established through the scoping review.

However, this study does not assess the external reliability of the questionnaire. Similar to Carolus et al. [15], we believe it is important to conduct further research to ensure that our instrument adequately measures the concept of AI literacy. Considering the limitations of PSE as a measure of competence or literacy, it would be appropriate to triangulate these results with data from a study evaluating students' performance on AI literacy tasks. Additionally, the questionnaire was used in a specific context with French-speaking undergraduate students enrolled in an undergraduate program in Quebec. Its use in a different context or language would require adaptation, for instance, through a cross-cultural validation process [30]. Finally, although several frameworks define AI literacy dimensions based on Bloom's revised taxonomy by Anderson and Krathwohl [e.g., 8, 10, 15], this questionnaire does not assess perceived self-efficacy at each taxonomic level, either globally or by AI literacy dimension. Instead, it evaluates the overall level of PSE in relation to AI literacy.

The next steps of this project will focus on analyzing the collected data to provide an overview of PSE in AI literacy among university student population. Statistical analyses (t-tests, ANOVA) will be conducted to identify significant differences among participants based on their gender, their membership in a First Nations community, their

citizenship status (local or international student), and their condition (physical or psychological disability). Analyses will also be conducted to identify factors influencing perceived PSE levels based on the field of study and proportion of completed courses in the program. Additionally, a cross-cultural validation process will be undertaken to adapt the questionnaire for teachers. This will enable future studies to assess both students' and teachers' AI literacy levels.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Pereira, A.C., Romero, F.: A review of the meanings and the implications of the Industry 4.0 concept. Procedia Manuf. **13**, 1206–1214 (2017)
2. World Economic Forum. The Future of Jobs Report 2023. 299 (2023). https://www.weforum.org/reports/the-future-of-jobs-report-2023/
3. Tremblay, C., Poellhuber, B.: Analyse qualitative de référentiels de compétences du XXIe siècle, numériques et informationnelles : tendances mondiales observées. Form. Prof. **30**, 1–26 (2022)
4. van Laar, E., van Deursen, A.J.A.M., van Dijk, J.A.G.M., de Haan, J.: The relation between 21st-century skills and digital skills: a systematic literature review. Comput. Hum. Behav. **72**, 577–588 (2017)
5. van Laar, E., van Deursen, A.J.A.M., van Dijk, J.A.G.M., de Haan, J.: Determinants of 21st-century skills and 21st-century digital skills for workers: a systematic literature review. SAGE Open **10**, 2158244019900176 (2020)
6. Ferrari, A., Punie, Y., Brečko, B.N.: DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe. Publications Office, Luxembourg (2013)
7. Vuorikari, R., Kluzer, S., Punie, Y.: DigComp 2.2, the digital competence framework for citizens: with new examples of knowledge, skills and attitudes. (Office des publications de l'Union européenne, Luxembourg, LU (2022)
8. Almatrafi, O., Johri, A., Lee, H.: A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Comput. Educ. Open **6**, 100173 (2024)
9. Chiu, T.K.F., Sanusi, I.T.: Define, foster, and assess student and teacher AI literacy and competency for all: Current status and future research direction. Comput. Educ. Open **7**, 100182 (2024)
10. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp 1–16. Association for Computing Machinery, New York (2020)
11. Miao, F., Shiohira, K.: AI Competency Framework for Students, vol. 80 (2024). https://unesdoc.unesco.org/ark:/48223/pf0000391105
12. Ministère de l'Éducation et de l'Enseignement supérieur du Québec (MEES). Cadre de référence de la compétence numérique, vol. 34 (2019). http://www.education.gouv.qc.ca/fileadmin/site_web/documents/ministere/Cadre-reference-competence-num.pdf
13. Venkatesh, V., Croteau, A.-M., Rabah, J.: Perceptions of effectiveness of instructional uses of technology in higher education in an era of Web 2.0. In: 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 110–119. IEEE (2014)

14. Saltos-Rivas, R., Novoa-Hernández, P., Rodríguez, R.S.: On the quality of quantitative instruments to measure digital competence in higher education: a systematic mapping study. PLoS ONE **16**, e0257344 (2021)
15. Carolus, A., Koch, M.J., Straka, S., Latoschik, M.E., Wienrich, C.: MAILS-Meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change-and meta-competencies. Comput. Hum. Behav. Artif. Hum. **1**, 1–10 (2023)
16. Knoth, N., et al.: Developing a holistic AI literacy assessment matrix – bridging generic, domain-specific, and ethical competencies. Comput. Educ. Open **6**, 100177 (2024)
17. Sillat, L.H., Tammets, K., Laanpere, M.: Digital competence assessment methods in higher education: a systematic literature review. Educ. Sci. **11**, 402 (2021)
18. Zhao, Y., Pinto Llorente, A.M., Sánchez Gómez, M.C.: Digital competence in higher education research: a systematic literature review. Comput. Educ. **168**, 104212 (2021)
19. Allen, B., McGough, A.S., Devlin, M.: Toward a framework for teaching artificial intelligence to a higher education audience. ACM Trans. Comput. Educ. **22** (2022)
20. Chiu, T.K.F., Ahmad, Z., Ismailov, M., Sanusi, I.T.: What are artificial intelligence literacy and competency? A comprehensive framework to support them. Comput. Educ. Open **6**, 100171 (2024)
21. Pinski, M., Benlian, A.: AI literacy for users – a comprehensive review and future research directions of learning methods, components, and effects. Comput. Hum. Behav. Artif. Hum. **2**, 1–22 (2024)
22. Bandura, A.: Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs (1986)
23. Hatlevik, O.E., Throndsen, I., Loi, M., Gudmundsdottir, G.B.: Students' ICT self-efficacy and computer and information literacy: determinants and relationships. Comput. Educ. **118**, 107–119 (2018)
24. Ulfert-Blank, A.-S., Schmidt, I.: Assessing digital self-efficacy: review and scale development. Comput. Educ. **191**, 104626 (2022)
25. Dunning, D.: Chapter five - the dunning–kruger effect: on being ignorant of one's own ignorance. In: Olson, J.M,. Zanna, M.P. (eds.) Advances in Experimental Social Psychology, vol. 44, pp. 247–296. Academic Press (2011)
26. Achim, A.: Esprit et enjeux de l'analyse factorielle exploratoire. TQMP **16**, 213–247 (2020)
27. Field, A.P.: Discovering Statistics Using IBM SPSS Statistics. Sage, Los Angeles (2018)
28. Béland, S., Cousineau, D., Loye, N.: Utiliser le coefficient oméga à la place de l'alpha de Cronbach. McGill J. Educ./Revue des sciences de l'éducation de McGill **52** (2017)
29. Béland, S., Michelot, F.: Une note sur le coefficient oméga (ω) et ses déclinaisons pour estimer la fidélité des scores. Mesure et évaluation en éducation **43**, 103–122 (2020)
30. Vallerand, R.J.: Vers une méthodologie de validation transculturelle de questionnaires psychologiques: Implications pour la recherche en langue française. Can. Psychol. **30**, 662 (1989)
31. Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ (2021). https://doi.org/10.1136/bmj.n71

ChatGPT's Impact on Students' Performance in Solving Tasks Requiring Critical Thinking

Kira Rohwer(✉) and Sabine Seufert

University of St. Gallen, St. Gallen, Switzerland
{kira.rohwer,sabine.seufert}@unisg.ch

Abstract. This study investigates the impact of ChatGPT-assistance compared to web search with Google on undergraduate students' task and learning performance in solving tasks that require critical thinking. A supervised online experiment with 39 undergraduate students from a Swiss business university was conducted. Participants completed 11 critical thinking tasks, deliberately designed to prevent ChatGPT from generating complete solutions in response to a single prompt, and a post-knowledge test to assess potential learning gains. The experimental group (EG, $n = 20$) used the generative AI tool ChatGPT-4.0 for assistance, while the control group (CG, $n = 19$) relied on conventional web search with Google. Welch's t-tests revealed no significant differences between groups in rubric-rated task performance ($p = .71$) or post-test scores ($p = .93$). Nevertheless, performance in analytical tasks approached statistical significance ($p = .06$), while surprisingly moderately favoring the Google-group ($d = 0.63$). These results challenge the assumption that AI tools like ChatGPT inherently enhance task performance. Instead, they indicate that independent critical thinking remains essential for achieving performance gains in complex tasks. The findings highlight the need for higher education curricula and task designs that actively integrate AI tools while preserving students' cognitive engagement, ensuring that they serve as a complement rather than a substitute for critical thinking.

Keywords: Generative AI · ChatGPT · Critical Thinking · Student Performance

1 Introduction

The rise of ChatGPT, a generative artificial intelligence (AI) tool, is reshaping higher education. Built on a large language model (LLM), ChatGPT provides students with personalized learning support and assists in concept clarification, information retrieval, and feedback processing [9]. However, its widespread adoption raises concerns that overreliance on AI may undermine students' ability to think critically [16].

Critical thinking (CT) is a fundamental skill in higher education [11] and essential for evaluating AI-generated outputs [10]. It involves analyzing information, assessing arguments, making reasoned decisions, and engagement in metacognition [6, 14] (see Sect. 2). Traditionally, students have used web search tools like Google to receive support to complete academic tasks that require these skills [17]. However, unlike Google, ChatGPT not only provides information but also generates structured responses, potentially

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 218–225, 2025.
https://doi.org/10.1007/978-3-031-99264-3_27

reducing cognitive effort [3]. While this may enhance task performance, it could also impede deeper learning processes and weaken independent CT [16]. Despite systematic reviews and empirical studies on ChatGPT's impact on CT [3, 4] and performance [1], research that investigates both aspects through the lens of task-solving remains limited [12]. Moreover, research comparing the effects of ChatGPT and Google on CT in higher education are still scarce [4], although the usage of both tools reflects students' real-world habits.

This study addresses the research gap by examining the impact of ChatGPT on undergraduate students' task performance in CT tasks and potential learning gains through a post-knowledge test. We address this by posing two research questions (RQs):

RQ1: Do undergraduate students using ChatGPT demonstrate a higher task performance (rubric-based assessment) in solving tasks requiring CT compared to those using Google web search?

RQ2: Do undergraduate students using ChatGPT demonstrate a higher knowledge test performance (post-multiple-choice-test) after solving tasks requiring CT compared to those using Google web search?

2 Related Work and Hypotheses

CT is a fundamental cognitive skill in education and personal life, involving purposeful judgment based on reasoning [5]. Theoretical models on CT highlight essential subdomains of CT. Halpern [6] emphasizes argument *analysis* and verbal *reasoning* in addition to hypothesis-testing and decision-making skills, while stressing the importance of *metacognition* for CT. In addition, Stein and Haynes [14] identify the *evaluation* of information as one key component of CT in higher education. CT skills are commonly assessed through multiple-choice (MC) tests [5] or problem-solving and written responses [6]. Research on ChatGPT's impact on CT is mixed but emerging. Some studies indicate a negative impact, with ChatGPT impeding CT in nursing education compared to traditional textbooks [13]. Other studies suggest that ChatGPT fosters CT awareness [2], enhances CT skills [4], and positively impacts overall higher-order thinking [3].

ChatGPT's effect on task performance suggests overall benefits. A meta-analysis [3] reported a positive effect on academic performance gains. In addition, several comparative studies revealed that ChatGPT supports students' performance and skills in complex tasks, such as robotics [1], research methodology [4], and creative problem-solving [15].

Based on the theoretical foundations of CT and its relevant subdomains as well as the emerging empirical evidence on ChatGPT's effects on CT and performance, we propose the following two hypotheses:

H1: *Task Performance*: The experimental group (EG; ChatGPT-assistance) achieves a higher overall task performance score, and higher task-domain scores (analysis, evaluation, reasoning, and metacognition), compared to the control group (CG; conventional Google web search).

H2: *Knowledge Test Performance*: The EG achieves a higher post-knowledge-test score compared to the CG.

3 Methodology

3.1 Participants

We targeted undergraduate students from a Swiss business university. We recruited 40 students through the university's obligatory academic writing course. Participation was voluntary and students received CHF 60 for participation. One student from the CG did not complete the study and was excluded. The final sample consisted of 39 students (EG: $n = 20$; CG: $n = 19$), largely in the second semester ($M = 2.31$, $SD = .73$). Students ranged in age from 18 to 23 years ($M = 20.40$, $SD = 1.23$), among 16 self-identified as female (% = 41.0), 22 as male (% = 56.4), and one as diverse (% = 2.6). Mother tongues were German (% = 64.1), French (% = 17.9), English (% = 17.9) and *Other* (% = 25.6).

3.2 Research Design and Procedure

Research Strategy. To answer both research questions, we employed an experimental approach. Participants were randomly assigned to one of two conditions: ChatGPT-4.0-assistance, no access to web search or other tools (EG), and conventional web search with Google, no access to ChatGPT or other tools (CG). ChatGPT was chosen as it represents a widely used generative AI tool. Google was selected as the control condition because it closely reflects students' real-world study habits. The experimental setting was selected to systematically compare how students process tasks requiring CT.

Experimental Design and Procedure. To assess our variables, we conducted a four-stage supervised online experiment between July and September 2024 via Qualtrics and Microsoft Teams in English. Each session lasted approximately 2.5 h. Students completed the survey independently and were required to keep their camera on. The experiment followed this chronological order for both conditions (EG and CG):

1. *Onboarding*: Participants received a brief introduction to the study's objectives and anonymity guidelines, followed by a supervised introduction to ChatGPT and Google web search to ensure tool familiarity. Students in the ChatGPT-group confirmed their exclusive use of ChatGPT, while those in the Google group agreed to solely use Google for web search and refrain from using any tools.
2. *Task-Set*: Students completed a task-set of 11 CT tasks under ChatGPT- or Google-use, assessing their performance (*Task Performance*). The supervisor provided a time check after one hour to support time management.
3. *Knowledge Test*: After completing the task-set, students took an eight-question MC post-knowledge test (*Knowledge Test Performance*), referring to AI and CT content addressed by the tasks. Before starting, students were required to close ChatGPT or Google and all open tabs to ensure no external resources were used.
4. *Demographic Data Collection*: Students provided demographic information.

In a post-treatment phase, two experts rated students' task performance between September and October 2024 with standardized rubrics. All responses were anonymized, securely stored, and processed for reliability measures and subsequent analyses.

3.3 Material and Data Processing

Task-Set and Rubrics. We developed a task-set with matching rubrics to measure *Task Performance*. Our task-set consists of 11 open-ended tasks designed to challenge students' CT in dealing with AI in an academic context [12]. Tasks were formed by classical CT frameworks and their domains [6, 14] (see Sect. 2), focusing on analysis, evaluation, reasoning, and metacognition. Other CT domains were not included, as the task-design aimed to stimulate students' CT engagement rather than to comprehensively assess it, as it would be the case with a standardized psychometric instrument. The tasks included both academic content relevant to students' field of study (i. e. economic graphs, student group work) and AI-related topics that students come across in their studies when using AI tools (i. e. AI-generated bias, reflection on ChatGPT-usage), ensuring alignment with students' experiences. In addition, the tasks incorporate on practices and results from assessment studies on generative AI [7, 8]. Moreover, tasks were crafted in a way that they could not be solved by ChatGPT without intermediate steps, i. e. by copying the task and receiving a perfect solution through a single prompt. To ensure this, each task was tested in ChatGPT 4.0. If the model provided a direct solution, the task was revised. This process step was implemented to guarantee that students in both groups were required to engage in CT on a holistic level. We structured our task-set into 11 tasks, allocated to four task domains:

- *Analysis*: Questioning underlying assumptions of AI-generated text (task 1), identifying and interpreting data (task 2), and detecting AI-generated biases (task 3).
- *Evaluation*: Evaluating arguments (task 4), interpreting and comparing AI-generated graphical outputs (task 5), and assessing AI-generated content (task 6).
- *Reasoning*: Applying logical processes to AI-generated information, including deductive reasoning and pattern recognition (task 7), inductive reasoning and conclusion drawing (task 8), and justifying decisions (task 9).
- *Metacognition*: Reflecting on general biases (task 10), and ChatGPT-usage in higher education (task 11), both encompassing self-awareness.

We developed 11 rubrics, specifically tailored for each task, assessing students' performance related to (1) reasoning quality, (2) completeness, and (3) content. Each rubric consisted of three criteria, ranging from 0 (insufficient) to 2 (excellent), totaling a maximum score of 6. For each participant, overall *Task Performance* was calculated as the average score per task. *Task Performance* within the CT domains was computed as the sum of the respective task scores, adjusted for the number of tasks per domain.

Multiple-Choice Test. We developed an eight-item MC-test to assess *Knowledge Test Performance*. Each question provided four response options with one correct answer. The test included exemplary questions on CT and AI to make assumptions about students' potential learning outcomes. Baseline knowledge was assumed to be comparable, as all were early-stage bachelor students. MC-questions' content was directly or indirectly addressed by the tasks beforehand. For each participant, *Knowledge Test Performance* was calculated as the average score (binary scale) per MC-question.

3.4 Reliability Measures

Pre-test. Prior to the experiment, we conducted a pre-test in May 2024 with five students of our target group, who tested the tasks under ChatGPT-assistance. Following the pre-test, the task-set and experimental procedure were refined and adapted.

Task Performance Ratings. Task responses were anonymously scored by two faculty researchers, experts in their fields. Inter-rater reliability between both raters was assessed by calculating weighted Cohen's cappa coefficient on the task score scale of 1 to 6 for all 11 tasks per student. Weighted Cohen's Cappa coefficient indicated a substantial agreement between both raters ($\kappa = 0{,}73$).

3.5 Data Analyses

Statistical Analyses. For RQ1 and RQ2, Welch's t-tests were performed to compare group means, standard deviations, and effect sizes (Cohen's *d*) to provide robust comparisons between EG and CG. To facilitate intuitive interpretation, overall *Task Performance* was also indicated as the average score per task in percentage. All statistical tests were performed with a standard confidence interval of 95%. The statistical tests were conducted using RStudio (version 4.4.1). The methods were chosen for insights into group differences and effect sizes while considering the small sample size.

4 Results

4.1 Using ChatGPT Does not Enhance Task Performance (RQ1)

To evaluate whether students of the ChatGPT-group (EG) demonstrated a higher performance in solving tasks targeting CT compared to the Google-web-search (CG), Welch's t-tests were conducted, and task performance scores (scale from 0 to 6) compared. In addition, the average task scores in percentage were calculated.

The results revealed no significant difference in overall task performance scores between the ChatGPT-group ($M = 3.34$, $SD = .80$) and the Google-group ($M = 3.25$, $SD = .80$), $t(36.91) = -.37$, $p = .71$, $d = -.12$, indicating a negligible effect. Additionally, overall task performance scores were nearly identical between the EG and the CG (EG: $\% = 55.67$; CG: $\% = 54.17$). Similarly, no significant differences were observed between both groups in task performance scores across the task domains of analysis, evaluation, decision-making, and metacognition ($p > .05$ for all comparisons). However, performance in the analysis task domain approached statistical significance ($p = .06$), with a moderate effect in favor of the CG ($d = .63$). Effect sizes for the evaluation, reasoning, and metacognition domains were small ($d = -.22$ to $-.43$), with effects favoring the EG.

In summary, H1 was not supported; nevertheless, the pattern of small effects favoring the ChatGPT-group leaves room for further investigation. The results of the t-test analyses and task performance scores are presented in Table 1 on the following page.

Table 1. Results of Welch's t-tests between Students in the ChatGPT-Group and the Google-Group on Task Performance.

Variable	ChatGPT-Group (EG)			Google-Group (CG)			t(df)	p	Cohen's d
	M	SD	%	M	SD	%			
Task Performance	3.34	0.80	55.67	3.25	0.80	54.17	$t(36.91) = -0.37$	0.71	−0.12
Analysis	2.59	0.95	43.17	3.18	0.94	53.00	$t(36.95) = 1.96$	0.06	0.63
Evaluation	3.78	1.08	63.00	3.29	1.16	54.83	$t(36.51) = -1.36$	0.19	−0.43
Reasoning	3.22	1.32	53.67	2.90	1.16	48.33	$t(36.50) = -0.80$	0.43	−0.26
Metacognition	4.01	1.06	66.83	3.80	0.88	63.33	$t(36.34) = -0.68$	0.50	−0.22

4.2 Using ChatGPT Does not Enhance Knowledge Test Performance (RQ2)

To evaluate whether students of the ChatGPT-group (EG) demonstrated a higher knowledge test performance compared to the Google-web-search (CG), Welch's t-tests were conducted and scores (binary scale) compared.

The results revealed no significant differences in the knowledge test performance scores between the ChatGPT-group ($M = .63$, $SD = .23$) and the Google-group ($M = .64$, $SD = .25$), $t(36.52) = .09$, $p = .93$, $d = .03$, indicating a negligible effect. Thus, H2 was not supported. The results of the t-test analyses are presented in Table 2 below.

Table 2. Results of Welch's t-tests between Students in the ChatGPT-Group and the Google-Group on Knowledge Test Performance.

Variable	ChatGPT-Group (EG)		Google-Group (CG)		t(df)	p	Cohen's d
	M	SD	M	SD			
Knowledge Test Performance	0.63	0.23	0.64	0.25	$t(36.52) = 0.09$	0.93	0.03

5 Discussion and Conclusion

This study examined whether ChatGPT impacts students' task performance (RQ1), and knowledge test performance (RQ2) compared to conventional web search with Google in the context of students completing tasks requiring CT in an experimental setting.

Regarding RQ1, the findings revealed that ChatGPT-assistance did not significantly enhance task performance compared to Google. No significant differences emerged in overall task performance or within specific CT task domains (analysis, evaluation, reasoning, and metacognition). However, while ChatGPT did not improve overall task performance, it also did not hinder it. This suggest that students may not yet know how to leverage ChatGPT effectively for task-solving. In addition, small effect sizes within evaluation, reasoning and metacognition tasks ($d = –.22$ to $–.43$) favored the ChatGPT-group. In contrast, task performance in the analysis domain showed a moderate effect favoring the Google-group ($d = .63$) and approached statistical significance ($p = .06$). This may suggest that students tend to over-rely on ChatGPT in tasks requiring analytical thinking, leading to lower critical engagement compared to those using Google.

With regard to RQ2, the findings revealed no significant improvement in the post-knowledge test, indicating no difference in learning gains between students using ChatGPT and Google. This result suggests that ChatGPT does not facilitate deeper learning. A possible explanation is that students may rely on ChatGPT for surface-level answers rather than engaging in CT-driven problem-solving.

Several limitations should be acknowledged. First, the small and specific sample limits generalizability, though the study offers initial insights into student performance in complex tasks. Second, CT was not measured using standardized psychometric tools; instead, tasks reflected academic reasoning demands. Third, task performance was assessed by two raters, with inter-rater reliability ensured through rubrics and calibration. Fourth, knowledge performance was measured without a pre-test; however, randomization likely balanced prior knowledge.

Future research should replicate these findings with larger and more diverse student populations to strengthen validity and investigate long-term effects on performance and CT. In addition, future work would benefit from the consideration of mediators, such as digital literacy or prior experience with AI tools. Educators should be trained to guide students in using ChatGPT and other tools based on LLMs to enhance, rather than replace, cognitive engagement. In our future work, we aim to refine and replicate the study in collaboration with an Asian university. We intend to improve and validate the design of our CT tasks, while exploring cultural influences. In addition, we plan on investigating pedagogical strategies to enhance CT skills using AI-based applications.

To conclude, our study demonstrates that simply introducing generative AI does not automatically enhance task performance or learning. However, using ChatGPT also does not decrease performance or learning gains in comparison to conventional tools like Google. The challenge lies in equipping students with the skills to effectively integrate generative AI into task-solving to leverage such tools – through CT.

Acknowledgments. We thank the 39 students who volunteered to participate in our study.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brender, J., El-Hamamsy, L., Mondada, F., Bumbacher, E.: Who's helping who? when students use ChatGPT to engage in practice lab sessions. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Lecture Notes in Computer Science, AIED 2024, vol. 14829, pp. 235–249. Springer, Cham (2024)
2. Darmawansah, D., Rachman, D., Febiyani, F., Hwang, G.-J.: ChatGPT-supported collaborative argumentation: Integrating collaboration script and argument mapping to enhance EFL students' argumentation skills. Educ. Inf. Technol. **30**, 3803–3827 (2025)
3. Deng, R., Jiang, M., Yu, X., Lu, Y., Liu, S: Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Comput. Educ. **227**, 105224 (2024)
4. Essel, H.B., Vlachopoulos, D., Essuman, A.B., Amankwa, J.O.: ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). Comput. Educ. Artif. Intell. **6**, 100198 (2024)
5. Facione, P.A.: Critical thinking: a statement of expert consensus for purposes of educational assessment and instruction. Research Findings and Recommendations. American Philosophical Association, Newark (1990)
6. Halpern, D.F.: Teaching critical thinking for transfer across domains: disposition, skills, structure training, and metacognitive monitoring. Am. Psychol. **53**(4), 449–455 (1998)
7. Hsiao, Y.-P., Klijn, N., Chiu, M.-S.: Developing a framework to re-design writing assignment assessment for the era of Large Language Models. Learn. Res. Pract. **9**(2), 148–158 (2023)
8. Huang, C.-W., Coleman, M., Gachago, D., Van Belle, J.-P.: Using ChatGPT to encourage critical AI literacy skills and for assessment in higher education. In: Van Rensburg, H.E., Snyman, D.P., Drevin, L., Drevin, G.R. (eds.) ICT EDUCATION. SACLA 2023, vol. 1862, pp. 105–118. Springer, Cham (2024)
9. Jahani, M., Baruah, B., Ward, A.: Exploring ChatGPT utilization among master's students in higher education. In: ITHET 2024, vol. 21, pp. 1–6. IEEE, Paris (2024)
10. Kenedy, R.A.: The challenges of critical thinking in the era of artificial intelligence. Eur. J. Multidisc. Stud. **9**(2), 19–36 (2024)
11. Liu, O.L., Frankel, L., Roohr, K.C.: Assessing critical thinking in higher education: current state and directions for next-generation assessment. ETS Res. Report Series **1**(2014), 1–24 (2014)
12. Seufert, S., Rohwer, K.: Developing a framework for analysing and assessing critical thinking skills for the responsible use of generative AI in higher education. In: Chova, L. G., Martínez, C. G., Lees, J. (eds.) 17th annual International Conference of Education, Research and Innovation, ICERI2024 Proceedings, vol. 17, pp. 8579–8585. IATED Academy, Valencia (2024)
13. Shin, H., De Gagne, J. C., Kim, S. S., Hong, M.: The impact of artificial intelligence-assisted learning on nursing students' ethical decision-making and clinical reasoning in pediatric care. CIN: Comput. Inf. Nurs. **42**(10), 704–711 (2024)
14. Stein, B., Haynes, A.: Engaging faculty in the assessment and improvement of students' critical thinking using the critical thinking assessment test. Change: Maga. High. Learn. **43**(2), 44–49 (2011)
15. Urban, M., et al.: ChatGPT improves creative problem-solving performance in university students: an experimental study. Comput. Educ. **215**(1), 10503 (2024)
16. Zhai, C., Wibowo, S., Li, L.D: The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn. Environ. **11**(28) (2024)
17. Zlatkin-Troitschanskaia, O., Hartig, J., Goldhammer, F., Krstev, J.: Students' online information use and learning progress in higher education – a critical literature review. Stud. High. Educ. **46**(10), 1996–2021 (2021)



Intent-Based vs. LLM-Based Chatbots in MOOCs: Do Diverse Students Experience Them Differently?

Songhee Han[1](✉), Chaewon Kim[1], and Hunhui Na[2]

[1] Florida State University, Tallahassee, FL 32311, USA
songhee.han@fsu.edu
[2] University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract. Despite the growing adoption of chatbots in massive open online courses (MOOCs), little research has examined how demographic factors shape students' learning experiences with different chatbot types. Existing studies focus on the general effectiveness of intent-based and large language model (LLM)-based chatbots, but how student characteristics influence their perceptions and use remains unclear. This study investigates how age, gender, socioeconomic region, and native language influence students' learning experiences with intent-based and LLM-based chatbots across five domains: social presence, teaching and cognitive presence, self-regulation, perceived ease of use, and behavioral intention to use. Survey data were collected from two MOOC courses, one utilizing an intent-based ($n = 340$) and another an LLM-based chatbot ($n = 121$). The results revealed that age significantly influenced students' learning experiences with the LLM-based chatbot, as younger students reported higher social presence, perceived ease of use, and behavioral intention to use it, whereas no significant age-related differences were found in the intent-based chatbot group. Socioeconomic status affected both chatbot users' social presence and teaching and cognitive presence, with lower-middle-income students reporting stronger levels in these learning domains across both chatbot groups. However, gender and native language had no significant effects in either chatbot group. These findings suggest that LLM-based chatbots may be more effective for younger students, while intent-based chatbots could support older students better. Additionally, the stronger presence perceptions among lower-middle-income students indicate the potential of chatbots to address educational disparities. Future research should explore hybrid chatbot models that integrate AI-driven flexibility with structured guidance to optimize learning experiences across diverse MOOC students.

Keywords: Demographics · Intent-based chatbot · LLM-based chatbot · MOOCs

1 Introduction

Massive open online courses (MOOCs) have transformed the educational landscape by providing accessible learning opportunities to a global audience [10]. Chatbots have been integrated into MOOCs to enhance student engagement and support. Until recently,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 226–233, 2025.
https://doi.org/10.1007/978-3-031-99264-3_28

intent-based chatbots have been the primary approach, operating by recognizing predefined user intents and responding accordingly [20]. These systems rely on structured frameworks to interpret user inputs and provide relevant responses. Conversely, large language model (LLM)-based chatbots have introduced more advanced conversational agents capable of understanding and generating human-like language, offering more contextually relevant interactions [20]. While previous studies have explored the effectiveness of these chatbots in education, limited research has explored how demographic factors shape students' learning experiences across various domains when interacting with one type versus another in MOOCs [21]. Understanding these influences is crucial for designing and implementing chatbots catering to diverse students.

This study aims to fill this gap by investigating how demographic factors—age, gender, region, and native language—affect five key learning experience domains: social presence, teaching and cognitive presence, self-regulation, perceived ease of use, and behavioral intention to use [8]. Specifically, it seeks to answer the research question: *How do demographic factors influence students' learning experience domains when using an intent-based chatbot versus an LLM-based chatbot in MOOCs?*

2 Relevant Previous Studies

Recent research highlights that while intent-based chatbots rely on predefined responses and structured interactions, LLM-based chatbots generate responses dynamically, allowing for more flexible and personalized communication [2, 23]. Additionally, demographic factors such as age, gender, socioeconomic status, and native language play a crucial role in shaping student engagement and learning outcomes in MOOCs, influencing the chatbot's effectiveness [5, 17]. Understanding the intersection of chatbot design and demographic influences is essential for optimizing their integration in MOOCs.

2.1 Intent-Based vs. LLM-Based Chatbots

Intent-based chatbots function within a structure, mapping user inputs to specific intents to trigger predefined responses [6]. This method ensures reliability and speed but may struggle with unexpected queries or nuanced conversations. These systems are also called rules-based chatbots, emphasizing their reliance on predefined rules and decision trees to guide interactions [18]. In contrast, LLM-based chatbots utilize extensive datasets to generate responses, enabling them to handle a broader range of topics and provide more engaging interactions. [1] also highlighted that AI chatbots, including LLM-based, offer personalized learning experiences, enhancing learning outcomes.

Further studies suggested that LLM-based chatbots significantly enhanced online courses' teaching and learning processes (e.g., [15]). However, despite their advantages, concerns regarding accuracy, bias, and computational demands remain key considerations when deploying LLMs in educational environments [22]. While LLM-based chatbots provide more dynamic and adaptive interactions, how they improve students' learning experiences may depend on individual characteristics like age and socioeconomic background. Thus, understanding how different demographics engage with intent-based and LLM chatbots can help optimize their effectiveness in MOOCs.

2.2 Demographic Factors' Influence in MOOCs and Trust in AI

Demographic factors significantly influence students' behaviors and outcomes in MOOCs, with variables such as age, gender, and educational background affecting course completion rates and engagement levels. Age plays a significant role, as younger students exhibit higher engagement levels, while older students demonstrate higher course completion rates [4]. Gender differences also influence MOOC participation; a study indicates men outnumber women in MOOCs [3]. Additionally, educational background and prior online learning experience influence MOOC success, with longer exposure to digital learning environments demonstrating higher completion rates [12]. A recent study highlighted disparities in AI adoption, with education level and digital literacy influencing usage more than income [9, 11]. These studies suggest that chatbot-assisted learning may help bridge digital gaps but requires effective strategies to support underserved students.

Socioeconomic status (SES) has also been found to shape students' engagement with AI-based tools. [13] argue that individuals from higher SES backgrounds exhibit greater trust in AI as they have more exposure to technology and perceive AI as valuable. This argument aligns with findings that higher SES students are generally more receptive to AI support, likely due to their prior experience with advanced technological tools. However, conflicting evidence exists. [16] found that SES plays a minimal role in determining trust in AI, with familiarity and technological competence being stronger predictors. Their findings suggest that students with greater technological experience, regardless of SES, are more likely to view AI as beneficial, while those with limited exposure may be more skeptical or resistant to AI-generated feedback.

3 Data Collection and Analysis

3.1 Study Context and Participants

This study examines students' learning experiences with two types of chatbots, intent-based and LLM-based, in MOOCs. Two survey datasets analyzed in this study were sourced from the same research site, a MOOC provider specializing in professional development courses for journalists in the Southwestern United States.

The intent-based chatbot dataset was collected from an election reporting course in 2022. A total of 3,358 students enrolled, and study invitations were sent to those confirmed as chatbot users based on chat logs to complete a Qualtrics survey on their perceptions and demographics. A total of 340 students submitted full survey responses. The chatbot was built using Dialogflow [7], relying on a pre-existing knowledge base.

The LLM-based chatbot dataset was obtained from a digital content creation course in 2024. This course had 4,918 students, and we collected 121 participants' completed survey responses using the same recruitment method described earlier per the intent-based chatbot data collection. The same Qualtrics survey structure was used to ensure consistency between the two datasets. The chatbot adopted Llama 3.2 [19] as a base model, integrating the same knowledge base as the intent-based but operating as a retrieval-augmented generation system [14].

English was the language of instruction in both courses, and the chatbot replaced the FAQ webpage component that was traditionally part of the course structure. The chatbot was integrated into every module; however, its use was voluntary.

3.2 Survey Data Collection and Analysis

The survey instrument for both data collection was a previously validated scale measuring students' learning experiences with chatbots in online courses [8]. The scale consisted of 22 items in five domains: social presence, teaching and cognitive presence, self-regulation, perceived ease of use, and behavioral intention to use. Responses were based on a five-point Likert scale, from 1 (completely disagree) to 5 (completely agree).

For the intent-based chatbot dataset, the age distribution was 18–24 ($n = 34$), 25–34 ($n = 91$), 35–44 ($n = 122$), 45–54 ($n = 59$), and 55 + ($n = 34$). More males than females participated (201 vs. 133, six disclosed), and most were non-native English speakers (274 vs. 66). For the LLM-based chatbot dataset, the age distribution was 18–24 ($n = 21$), 25–34 ($n = 29$), 35–44 ($n = 26$), 45–54 ($n = 32$), and 55 + ($n = 13$). More males than females participated (67 vs. 53, one non-binary), and most were non-native English speakers (107 vs. 14). Table 1 presents the socioeconomic representations.

For both chatbot datasets, Shapiro-Wilk normality tests indicated that the distribution of all demographic variables was non-normal ($p < .001$). Similarly, both groups' overall learning experience scores deviated from normality ($p < .001$, $p = .003$). Therefore, non-parametric statistical tests were conducted. Post-hoc analyses were performed for significant Kruskal-Wallis results to determine difference details.

The results will examine these factors' influences on learning experiences with the LLM-based, highlighting both consistent and divergent findings across chatbot types.

Table 1. Distribution of students based on socioeconomic regions and chatbot types

Socioeconomic region	Intent-based chatbot users				LLM-based chatbot users			
	low	Lower-middle	Upper-middle	high	low	Lower-middle	Upper-middle	high
Student count	52	233	28	27	7	62	28	24
Total count	340				121			

4 Results

4.1 Reliability of the Learning Experience Scale

The scale's internal consistency demonstrated strong reliability across all factors for both chatbots (Cronbach's $\alpha > 0.80$) except for behavioral intention (0.77, 0.78). Despite this slight deviation, overall internal consistency remained excellent (0.95, 0.94).

4.2 Demographic Factors' Influences in the LLM-Based Chatbot Experience

The analysis of the LLM-based dataset revealed that gender and native language did not significantly influence students' learning experiences, which aligns with the findings from the intent-based chatbot dataset. However, age had a notable effect on learning experiences in the LLM-based environment, whereas the intent-based chatbot dataset showed no significant age-related differences. A Kruskal-Wallis test on age factor in the LLM-based dataset revealed significant differences across age groups in social presence ($p = .026$), perceived ease of use ($p = .007$), and behavioral intention ($p = .050$).

Post-hoc analysis indicated that students aged 18–24 scored significantly higher than those 45–54 in social presence ($p = .003$), behavioral intention ($p = .003$), and overall learning experience ($p = .005$). Regarding perceived ease of use, students 18–24 also scored significantly higher than those 25–34 ($p = .001$), 35–44 ($p = .004$), and 55 + ($p = .001$). These results indicate that younger students found the LLM-based chatbot more engaging and easier to use than older students (Fig. 1).

Socioeconomic region also played a significant role in students' learning experiences with the LLM-based chatbot, a pattern consistent with the intent-based chatbot dataset with varying degrees. A Kruskal-Wallis test showed statistically significant differences in social presence ($p < .001$) and teaching and cognitive presence ($p = .035$). Post-hoc

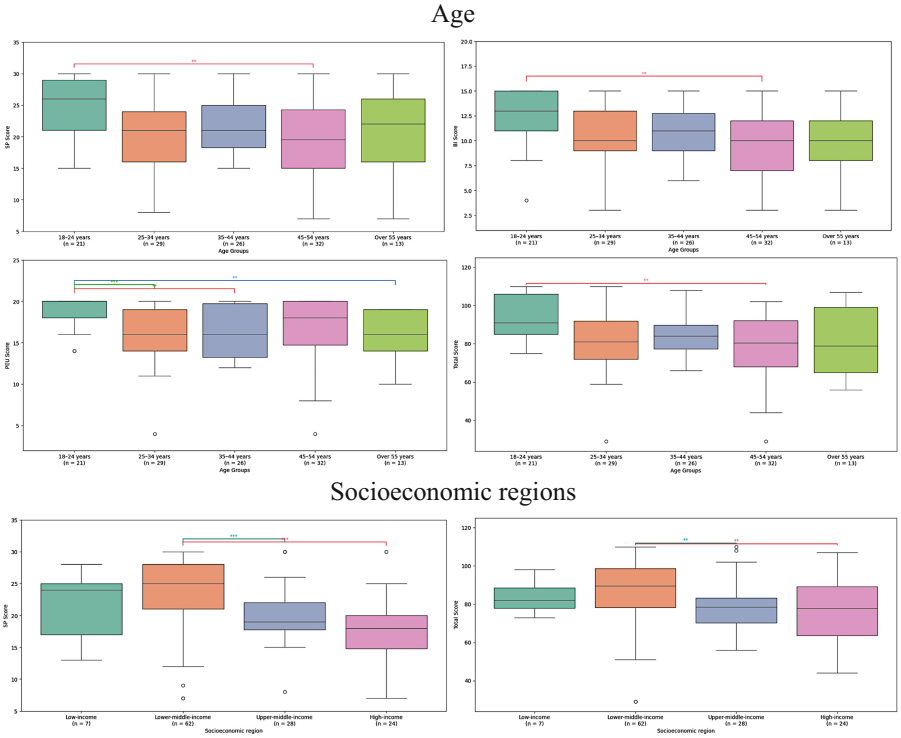


Fig. 1. Box plots of post-hoc analysis. BI (behavioral intention), SP (social presence), PEU (perceived ease of use), Total Score (overall learning experience). ***$p < .001$ and **$p < .01$.

analysis indicated that students from lower-middle-income region exhibited significantly higher social presence scores compared to those from upper-middle-income ($p < .001$) and high-income regions ($p < .001$). A significant difference was also observed in overall learning experience scores between students from lower-middle-income and those from upper-middle-income ($p = .008$) and high-income ($p = .005$) regions (Fig. 1).

5 Discussion

The findings suggest that younger students and students from lower-middle-income region engage more actively with the LLM-based chatbot, while students from higher-income regions and older age groups may have different expectations or challenges. The discussion will explore these patterns further by examining age-related differences and how socioeconomic regions influence students' chatbot perceptions.

5.1 Age Differences in Chatbot Engagement

The findings indicate that younger students had more favorable experiences with the LLM-based chatbot but not with the intent-based chatbot, suggesting that chatbot types may shape age-related differences in user experience. These results align with prior research suggesting that younger students are more comfortable with AI-driven conversational agents due to their greater digital literacy and familiarity [5, 9].

The absence of age-related differences in the intent-based chatbot dataset suggests that structured chatbot interactions provide a more uniform experience across age groups. Unlike LLM-based chatbots, which generate open-ended responses, intent-based chatbots rely on predefined structures that may be easier for older students to navigate. Previous research has indicated that older students often struggle with AI-generated responses that require contextual interpretation [17]. These findings suggest that LLM-based chatbots offer younger students deeper engagement but may introduce usability challenges for older students.

5.2 Socioeconomic Region and Chatbot-Assisted Learning Experience

The study found that socioeconomic regions significantly influenced the learning experience with chatbots, though the patterns varied between types. In both cases, students from lower-middle-income region reported stronger social presence and teaching and cognitive presence than their upper-middle- and high-income peers. These findings align with [11] suggesting that students in areas with fewer educational resources tend to rely more on digital tools to supplement their learning.

However, in the LLM-based chatbot dataset, both upper-middle- and high-income regions showed lower social, teaching and cognitive presence than lower-middle-income regions, indicating a broader socioeconomic influence than the other, where only low- and upper-middle-income differences were significant. This observation contrasts with [13], which argues people from higher socioeconomic backgrounds trust AI more, likely due to increased exposure to advanced tools. Conversely, [16] found that socioeconomic status played a minimal role in AI trust, with prior technological experience being a

more significant predictor, which was more aligned with this study's findings. These perspectives suggest that while AI access may be higher in wealthier regions, learning experiences with it may also be shaped by familiarity and expectations.

6 Conclusion

The findings highlight the need for chatbot implementations that balance structured support (strength of intent-based chatbot) with AI-driven adaptability (strength of LLM-based chatbot) to accommodate diverse student demographics. Future research should explore optimizing chatbot interactions by tailoring responses to users' prior AI exposure and learning needs.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Albadarin, Y., Saqr, M., Pope, N., Tukiainen, M.: A systematic literature review of empirical research on ChatGPT in education. Disc. Educ. **3**(1), 60 (2024). https://doi.org/10.1007/s44217-024-00138-2
2. Alsafari, B., Atwell, E., Walker, A., Callaghan, M.: Towards effective teaching assistants: from intent-based chatbots to LLM-powered teaching assistants. Nat. Lang. Process. J. **8**, 100101 (2024). https://doi.org/10.1016/j.nlp.2024.100101
3. Bayeck, R.Y., Hristova, A., Jablokow, K.W., Bonafini, F.: Exploring the relevance of single-gender group formation: what we learn from a massive open online course (MOOC). Br. J. Edu. Technol. **49**(1), 88–100 (2018). https://doi.org/10.1111/bjet.12528
4. Cagiltay, N. E., Toker, S., Cagiltay, K.: Exploring MOOC students' behavioural patterns considering age, gender and number of course enrolments: insights for improving educational opportunities. Open Praxis **16**(1) (2024). https://doi.org/10.55982/openpraxis.16.1.543
5. Cheng, X., Qiao, L., Yang, B., Li, Z.: An investigation on the influencing factors of elderly people's intention to use financial AI customer service. Internet Res. **34**(3), 690–717 (2024). https://doi.org/10.1108/INTR-06-2022-0402
6. Chien, Y.H., Yao, C.K.: Development of an AI userbot for engineering design education using an intent and flow combined framework. Appl. Sci. **10**(22), 7970 (2020). https://doi.org/10.3390/app10227970
7. Google Cloud. https://cloud.google.com/dialogflow/docs. Accessed 14 Feb 2025
8. Han, S., Hamilton, X., Cai, Y., Shao, P., Liu, M.: Knowledge-based chatbots: a scale measuring students' learning experiences in massive open online courses. Educ. Tech. Res. Dev. **71**(6), 2431–2456 (2023). https://doi.org/10.1007/s11423-023-10280-7
9. Jin, K.Y., Reichert, F., Cagasan, L.P., Jr., de La Torre, J., Law, N.: Measuring digital literacy across three age cohorts: exploring test dimensionality and performance differences. Comput. Educ. **157**, 103968 (2020). https://doi.org/10.1016/j.compedu.2020.103968
10. Johnston, J.S., et al.: Global use, adaptation, and sharing of massive open online courses for emergency health on the OpenWHO platform: survey study. J. Med. Internet Res. **27**, e52591 (2025). https://doi.org/10.2196/52591
11. Kacperski, C., Ulloa, R., Bonnay, D., Kulshrestha, J., Selb, P., Spitz, A.: Characteristics of ChatGPT users from Germany: implications for the digital divide from web tracking data. PLoS ONE **20**(1), e0309047 (2025). https://doi.org/10.1371/journal.pone.0309047

12. Kizilcec, R. F., Piech, C., & Schneider, E.: Deconstructing disengagement: analyzing student subpopulations in massive open online courses. In: Proceedings of the third international conference on learning analytics and knowledge, (2013, April)
13. Kozak, J., Fel, S.: How sociodemographic factors relate to trust in artificial intelligence among students in Poland and the United Kingdom. Sci. Rep. **14**(1), 28776 (2024). https://doi.org/10.1038/s41598-024-80305-5
14. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Process. Syst. **33**, 9459–9474 (2020)
15. Neumann, A.T., Yin, Y., Sowe, S., Decker, S., Jarke, M.: An LLM-driven chatbot in higher education for databases and information systems. IEEE Trans. Educ. **68**(1), 103–116 (2025). https://doi.org/10.1109/TE.2024.3467912
16. Novozhilova, E., Mays, K., Paik, S., Katz, J.E.: More capable, less benevolent: trust perceptions of AI systems across societal contexts. Mach. Learn. Knowl. Extract. **6**(1), 1 (2024). https://doi.org/10.3390/make6010017
17. Spulber, D., Guido Franco Amoretti, Siri, A.: The use of AI for education in third age: The role of EU project. https://www.unipegaso.iris.cineca.it/handle/20.500.12607/45203. Accessed 15 Feb 2024
18. Thorat, S. A., Jadhav, V.: A review on implementation issues of rule-based chatbot systems. In: Proceedings of the International Conference on Innovative Computing & Communications (ICICC) (2020). https://doi.org/10.2139/ssrn.3567047
19. Touvron, H., et al.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
20. Wyawahare, M., Roy, S., Zanwar, S.: Generative vs intent-based chatbot for judicial advice. In: 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, pp. 1–6. IEEE (2024). https://doi.org/10.1109/IATMSI60426.2024.10502550
21. Xue, Y., Khalid, F. B., Karim, A. B. A.: Emerging trends in self-regulated learning: a bibliometric analysis of MOOCs and AI-enhanced online learning (2014–2024). Int. J. Learn. Teach. Educ. Res. **24**(1), 420–442 (2025). https://doi.org/10.26803/ijlter.24.1.21
22. Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S., Ozcan, A.: Large language model-based chatbots in higher education. In: Advanced Intelligent Systems, p. 2400429 (2024)
23. Zobel, T., Meinel, C.: Comparing AI in online learning: the transition and trade-offs between intent-based learning assistants and LLM-chatbots in MOOCs. In: 2024 IEEE Digital Education and MOOCS Conference, pp. 1–6 (2024). https://doi.org/10.1109/DEMOcon63027.2024.10748211

A GenAI-Enhanced Platform for Personalized Doctoral Progress and Well-Being Support Through Single-Case Analytics

Mohamed Saban(✉), Luis P. Prieto, and Yannis Dimitriadis

GSIC-EMIC Research Group, Universidad de Valladolid, Valladolid, Spain
{mohamed.saban,luispablo.prieto}@uva.es, yannis@tel.uva.es

Abstract. Doctoral education faces high attrition rates and widespread well-being challenges. Research has identified key motivational factors distinguishing persisting students from those who withdraw: perceived progress, exhaustion levels, and thesis topic appropriation. Despite this knowledge, technological interventions, especially AI-based ones, targeting doctoral education remain scarce, partly due to each doctoral journey's uniqueness. This paper presents the co-design process and initial evaluations of a novel generative AI (GenAI)-enhanced system supporting doctoral students' perception of progress and well-being. The system facilitates practices (e.g., journaling, reflection) targeting the three aforementioned motivational factors that doctoral training interventions suggest can to improve student well-being. The system's key innovation is its combination of machine learning-based open learner models (based on individual's journaling) with GenAI explanations, enabling personalized student support for the analysis, visualization and sense-making of their progress and their specific predictive factors. Initial evaluations with full-time (N = 8) and part-time (N = 10) doctoral student cohorts during 4-week training periods show high usability and likelihood-to-recommend. Feedback also reveals improvement areas and design tensions, particularly regarding strategies for consistent platform engagement.

Keywords: Doctoral education · Single-case learning analytics · Generative AI · Well-being · Socio-emotional learning · Human-AI Collaboration

1 Introduction

Doctoral education research has shown that a high prevalence of mental health issues [22] among doctoral students, as well as high dropout rates [26]. Despite accumulated knowledge on the structural, social, and individual factors that contribute to these intertwined problems, effective evidence-based interventions to address them remain scarce [13,15].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 234–241, 2025.
https://doi.org/10.1007/978-3-031-99264-3_29

The inherent uniqueness of the doctoral experience, shaped by a distinct thesis topic, research process, existing skillset, as well as social and personal context of the student, makes one-size-fits-all solutions to these problems likely to fail [17]. While learning analytics (LA) and AI have gained traction elsewhere in education, their application in doctoral support is limited (see Di Mitri et al.'s Learning Pulse system to understand doctoral students' experiences of flow [8] for a notable exception). Yet, AI's capabilities to flexibly personalize learning experiences apparently matches well the uniqueness-driven problems of doctoral persistence and well-being.

In this regard, a promising strand of research has uncovered key motivational factors that students can influence, such as perceived progress, exhaustion, and appropriation of one's thesis work [5,7]. These factors, based upon self-determination theory [6], point to the learning of certain socio-emotional skills as a way forward for doctoral students [19]. Indeed, doctoral training targeting these factors through specific practices (journaling, structured reflection) have shown promising benefits for their well-being [16].

This work presents a novel AI-augmented system to support doctoral student progress, well-being, and topic appropriation by supporting the aforementioned practices and providing statistical machine learning open learner models (OLM) [3] of progress and analytics-driven, tailored recommendations and insights derived using GenAI.

2 Co-designing Progress and Well-Being Support with Doctoral Students

This study is part of a broader design-based research (DBR) [25] initiative to support doctoral student well-being and persistence using learning analytics (LA) and AI. Recognizing the inherently ethical implications of educational technology, especially when targeting student well-being, we have adopted a Value-Sensitive Design (VSD) approach [10], which is still rare in the field of AIED, VSD integrates systematically human values into technology design. Several iterations of co-design have been conducted with doctoral students from multiple disciplines, different profiles (part-time, full-time) in two countries (Spain and Estonia). In this process we combined methods from VSD (e.g., value scenarios) [10], values research in social and cross-cultural psychology [20], and mixed-method inquiry processes we have tailored for educational technology contexts.

Our first two co-design iterations [18], used questionnaires, diaries, and interviews with $N = 19$ doctoral students from diverse disciplines in the two countries. Here, self-direction emerged as a key value for students, along with the importance of perceiving progress – although this perception was found to be highly individualized due to varied contextual factors. Participants also highlighted struggles with focus and time tensions between thesis and non-thesis activities as major barriers to progress.

A subsequent co-design iteration with $N = 10$ part-time doctoral students (a collective especially vulnerable to dropout and low well-being) reaffirmed the

importance of self-direction and the value of learning about one's thesis topic [18]. Across diverse profiles, the most prominent technology-related value was usefulness, defined as the system's capacity to augment their capabilities, while fragmentation of tasks, time, and attention emerged as a major challenge to sustaining progress. The recent rise of GenAI influenced this latter co-design iteration, with students envisioning AI to support focus, reduce fragmentation, and help them perceive progress within their limited time and attention.

3 Doctoral Educational Technology (DET)

Research highlights that doctoral students often face motivational challenges, with perceived progress being critical to sustaining engagement [5,7]. Practices like journaling and structured reflection have shown promise in supporting such perception of progress and well-being [2,16]. To support doctoral students in this direction, we developed the Doctoral Educational Technology (DET) web-based platform. DET adopts a human-AI collaboration approach based on individualized analytics, similar to what some authors have called idiographic or single-case learning analytics [17,21]. In DET, doctoral students actively engage in data gathering through such progress-promoting practices. Each individual's data are then analyzed (using statistical machine learning and GenAI-based natural language processing) to build multiple interpretable learner models that are made available to students (hence, OLMs [3]) through visualizations and GenAI-based explanations. This, in turn, can foster the improvement of socioemotional skills such as openness to experience, valued action, and behavioral awareness, according to modern evidence-based therapy approaches such as Acceptance and Commitment Therapy (ACT) [9].

DET allows doctoral students to collect individualized qualitative and quantitative data (daily self-tracking, journaling, periodic reflections about progress) through customizable diary forms, administered via a Laravel-based interface. After data cleaning and chronological organization, student-provided data is pre-processed using keyword-based (to map personalized questions to relevant constructs) and large language model (LLM)-based extraction (to detect particular contextual elements from the narrative journaling of a day's events, using the ACAD framework [11]). Multiple learner models of progress (e.g., predict that student's satisfaction with progress as a function of different quantitative/qualitative aspects of that day) are then built (and updated), using simple, interpretable statistical machine learning techniques. We currently use LASSO regression [24] and regression trees for their transparency and interpretability in low-sample, personalized contexts, but other techniques could be added as more data points are gathered. These OLMs (along with other descriptive views of the student's data) are shown to students via Canva charts [4]. Given the breadth of disciplines (and, consequently, data and visualization literacy) of doctoral students, and to foster interpretability and actionability, explanations with reflective questions are dynamically generated using the OLM/visualization data, by an LLM (we are currently using Meta's LLaMa 3.2 model [1]). These prompts were

designed iteratively during co-design sessions to align with students' vocabulary and context. Figure 1 represents the general architecture of the DET platform.

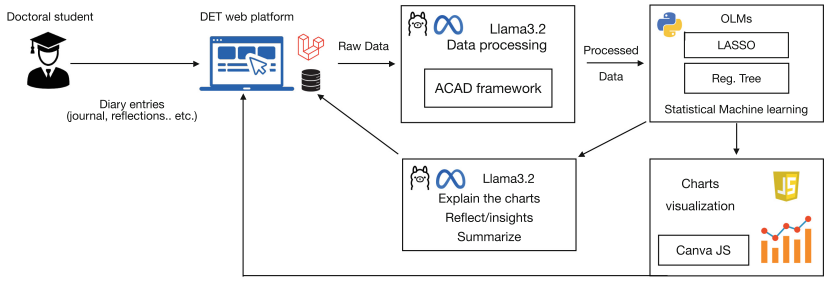
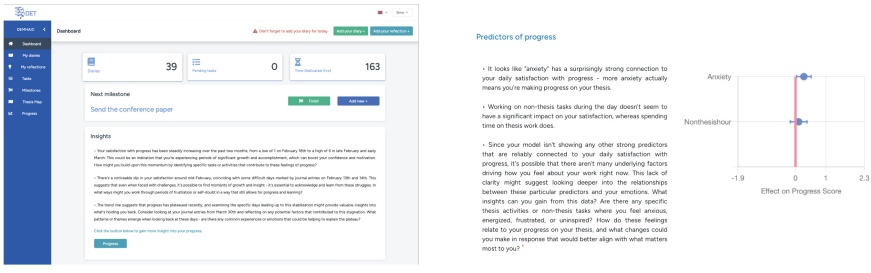


Fig. 1. Overview of the DET platform architecture: diary data is processed via the ACAD framework, modeled with ML techniques, and visualized with AI-generated insights from Llama 3.2.

Upon accessing the DET platform, doctoral students are presented with a personalized dashboard summarizing basic progress indicators and an LLM-generated summary of the insights and reflections derived from the OLMs/visualizations (Fig. 2a. To foster students' thesis appropriation, they are encouraged to develop a Thesis Map, which outlines key milestones and obstacles. Aside from the aforementioned entry of daily journals (including a self-assessment of progress satisfaction), students are also encouraged to periodically reflect on their progress, by accessing a progress dashboard. Figure 2b illustrate a coefficient plot from a LASSO regression model quantifying the impact of two predictors with the AI-generated explanations.



(a) (b)

Fig. 2. Screenshots from DET: (a) Main dashboard showing AI-generated insights. (b) Coefficient plot generated using LASSO regression, illustrating predictors of progress based on AI-generated insights.

4 Initial Evaluation

Context. In the current iteration of our DBR, we conducted preliminary usability and acceptance evaluations of the DET system among doctoral students. These evaluations were embedded into two authentic doctoral training interventions at a medium-sized university in Spain that followed the progress-oriented workshop structure described in [18]. Each intervention comprised four weekly 2.5-hour synchronous sessions with optional asynchronous activities (including DET platform use during these weeks). The training highlighted PhD persistence and well-being challenges and the significance of motivational factors like progress, well-being and appropriation [7] (and practices to enhance them). Participants came from all disciplines, with Training 1 delivered face-to-face primarily to on-site full-time doctoral students, and Training 2 delivered online for part-time or distance students.

Methods. To assess platform usability and acceptance while guiding further development, we employed a convergent parallel mixed methods research design (triangulating quantitative and qualitative data to enhance validity), iteratively for each of the two trainings. During the first session, participants were guided through DET's onboarding process and were encouraged to use key features (journal, thesis map, progress visualizations/explanations and periodic reflection) throughout the training. After three weeks of platform usage (primarily journaling), participants accessed learner model visualizations showing progress satisfaction and potential predictors (in the fourth session). Training 1 provided access to learner model visualizations *without* AI-generated explanations, while Training 2 included visualizations *with* AI-generated explanations.

To evaluate these socio-technical interventions, participants completed pre/post questionnaires measuring emotional well-being and socio-emotional skills, as well as platform evaluation questions. Our reporting here focuses on post-questionnaire responses to platform evaluation questions. This included standardized instruments like UMUX-LITE [14] for usability and Net Promoter Score (NPS) for likelihood-to-recommend. In Training 2 we also adapted six items from the EFLA instrument [23] to understand how the visualization of the learner models (and AI-generated explanations) were perceived by part-time students. Participants provided qualitative feedback through open-ended prompts (e.g., "I like," "I wish," "What if" [12]). To encourage honest input on the platform and well-being, data collection was anonymous, using non-identifiable nicknames. Quantitative data was analyzed descriptively, and open responses were inductively coded by the second author. This design balanced authenticity of data, practical constraints and honesty (through anonymous participation), in order to support rapid feedback for iterative platform improvement. However, this also limits the generalizability and internal validity of the results.

Available Evaluation Data. From the initial student pool who provided anonymous data at the outset of the trainings (N = 25 and N = 16 for Training 1 and 2), only a subset (N = 8 and N = 10, respectively) completed the final

questionnaire with DET evaluation data. Furthermore, in both trainings only 15 students entered at least one journal entry (with approximately half of them doing only 1–2 entries over the length of the course). This participation drop is common in authentic doctoral trainings with anonymous and voluntary data collection that is not tied to the training's assessment (as the primary aim of trainings was student support, not tool evaluation).

Results – Descriptive Statistics. All Training 1 respondents (N = 8) would recommend the platform to other doctoral students, scoring 9+ on the Net Promoter Score (NPS) scale. Usability received positive ratings (mean = 6.875 on a 1–7 scale), with system capabilities meeting student needs (mean = 6.375 and no rating below 6). The second training for part-time students (N = 10) yielded less positive results, with only 2 "promoters" (rating 9–10) and average NPS rating of 7.44. Ease of use remained highly rated (mean = 6.5) but not all respondents felt their requirements were met (mean = 5.7). This group's responses to EFLA questions about visualizations with GenAI explanations indicated positive views of SCLA models and explanations (overall mean = 5.93, std = 0.38 in a 1–7 Likert scale). Students valued the highest that the system encouraged them to change behavior and the awareness it provided of their current situation (mean = 6.22 and 6.11). The question about the system improving motivation for learning (thesis work) received positive but more heterogeneous scores (mean = 5.44, std = 1.67).

Results – Qualitative Analysis. Open responses supported the quantitative results: participants praised the platform's ease of use (e.g., *"user-friendly interface, ease of use"* (T1P5)) and simple progress visuals (e.g., *"Seeing the hours go up is comforting"* (T1P4)). Part-time doctoral students valued how it structured their thesis work (e.g., *"structured monitoring"* (T2P2); *"organizes information for reflection and planning"* (T2P3)), echoing concerns about fragmentation (Sect. 2). Some lower scores were linked to anxiety from unmet goals (e.g., *"I feel anxious if I haven't met the objective"* (T2P9)) and doubts about long-term usefulness. Suggested improvements included a mobile app (especially for reminders), personalized tracking, AI-based recommendations, and social features (e.g., *"AI suggestions for research planning and time management"* (T1P5)).

5 Conclusion

In response to dropout and mental health challenges in doctoral education, this paper presented DET, a digital platform combining analytics-based OLMs and GenAI to support doctoral progress, well-being, and persistence. Contributions include value-sensitive participatory design process and a novel integration of analytics-driven OLMs with LLMs to enhance reflection and actionability. Initial evaluations show promising usability, though low uptake suggests the need to better communicate value early on. Limitations include the exclusion of supervisors in co-design and short-term evaluation. Future work will refine progress analytics OLMs, personalized feedback, and GenAI-supported reflection tools like

the Thesis Map. We hope this work stimulates broader discussion on designing AIED systems that integrate AI with evidence-based training and motivational scaffolds to support human values and socio-emotional learning.

Acknowledgments. This work has been co-funded by grants PID2023-146692OB-C32 and RYC2021-032273-I, all financed by MCIN/ AEI/ 10.13039/501100011033. These grants have been co-funded by the European Union's ERDF and NextGenerationEU/ PRTR. The present work has also been supported by the Regional Government of Castile and Leon and FEDER, under project grant VA176P23.

References

1. AI, M.: Llama 3.2: Compact multilingual large language models for diverse applications (2024). https://arxiv.org/abs/2310.12346. Accessed 02 Feb 2025
2. Amabile, T., Kramer, S.: The Progress Principle: Using Small Wins to Ignite Joy, Engagement, and Creativity at Work. Harvard Business Press (2011)
3. Bull, S., Kay, J.: Open learner models. In: Advances in Intelligent Tutoring Systems, pp. 301–322. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14363-2_15
4. Canva: About canva (2025). https://www.canva.com/about/. Accessed 02 Feb 2025
5. De Clercq, M., Frenay, M., Azzi, A., Klein, O., Galand, B.: All you need is self-determination: investigation of phd students' motivation profiles and their impact on the doctoral completion process. Int. J. Doct. Stud. **16**, 189–209 (2021). https://doi.org/10.28945/4702
6. Deci, E.L., Ryan, R.M.: Self-determination theory: a macrotheory of human motivation, development, and health. Can. Psychol./Psychologie canadienne **49**(3), 182–185 (2008). https://doi.org/10.1037/a0012801
7. Devos, C., Boudrenghien, G., Nicolas Van der Linden, Azzi, A., Frenay, M., Galand, B., Klein, O.: Doctoral students' experiences leading to completion or attrition: a matter of sense, progress and distress. Eur. J. Psychol. Educ. **32**(1), 61–77 (2017)
8. Di Mitri, D., Scheffel, M., Drachsler, H., Börner, D., Ternier, S., Specht, M.: Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 188–197. ACM (2017)
9. Francis, A.W., Dawson, D.L., Golijani-Moghaddam, N.: The development and validation of the Comprehensive assessment of Acceptance and Commitment Therapy processes (CompACT). J. Contextual Behav. Sci. **5**(3), 134–145 (2016). https://doi.org/10.1016/j.jcbs.2016.05.003
10. Friedman, B., Hendry, D.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press, Cambridge (2019)
11. Goodyear, P., Carvalho, L., Yeoman, P.: Activity-centred analysis and design (ACAD): core purposes, distinctive qualities and current developments. Educ. Tech. Res. Dev. **69**(2), 445–464 (2021). https://doi.org/10.1007/s11423-020-09926-7
12. Hasso Plattner Institute of Design at Stanford University: Design Thinking Bootleg (2025). https://dschool.stanford.edu/tools/design-thinking-bootleg

13. Jackman, P.C., Jacobs, L., Hawkins, R.M., Sisson, K.: Mental health and psychological wellbeing in the early stages of doctoral study: a systematic review. Eur. J. High. Educ. 1–21 (2021). https://doi.org/10.1080/21568235.2021.1939752
14. Lewis, J.R., Utesch, B.S., Maher, D.E.: UMUX-LITE: when there's no time for the SUS. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2099–2102. ACM, Paris (2013). https://doi.org/10.1145/2470654.2481287
15. Mackie, S.A., Bates, G.W.: Contribution of the doctoral education environment to PhD candidates' mental health problems: a scoping review. High. Educ. Res. Dev. **38**(3), 565–578 (2019)
16. Prieto, L., Odriozola-González, P., Rodríguez-Triana, M.J., Dimitriadis, Y., Ley, T.: Progress-oriented workshops for doctoral well-being: evidence from a two-country design-based research. Int. J. Doct. Stud. **17**, 039–066 (2022). https://doi.org/10.28945/4898
17. Prieto, L.P., Pishtari, G., Dimitriadis, Y., Rodríguez-Triana, M.J., Ley, T., Odriozola-González, P.: Single-case learning analytics: feasibility of a human-centered analytics approach to support doctoral education. JUCS - J. Univ. Comput. Sci. **29**(9), 1033–1068 (2023). https://doi.org/10.3897/jucs.94067
18. Prieto, L.P., Rodríguez-Triana, M.J., Dimitriadis, Y., Pishtari, G., Odriozola-González, P.: Designing technology for doctoral persistence and well-being: findings from a two-country value-sensitive inquiry into student progress. In: Viberg, O., Jivet, I., Muñoz-Merino, P.J., Perifanou, M., Papathoma, T. (eds.) Responsive and Sustainable Educational Futures, vol. 14200, pp. 356–370. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42682-7_24
19. Prieto, L.P., Rodríguez-Triana, M.J., Odriozola-González, P., Dimitriadis, Y.: Single-case learning analytics to support social-emotional learning: the case of doctoral education. In: Wang, Y.E., Joksimović, S., San Pedro, M.O.Z., Way, J.D., Whitmer, J. (eds.) Social and Emotional Learning and Complex Skills Assessment, pp. 251–278. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-06333-6_12
20. Sagiv, L., Schwartz, S.H.: Personal values across cultures. Annu. Rev. Psychol. **73**(1), 517–546 (2022). https://doi.org/10.1146/annurev-psych-020821-125100
21. Saqr, M., Lopez-Pernas, S.: Idiographic learning analytics: a single student (n = 1) approach using psychological networks. In: Proceedings of the NetSciLA21 Workshop, vol. 2868, pp. 16–22 (2021)
22. Satinsky, E.N., et al.: Systematic review and meta-analysis of depression, anxiety, and suicidal ideation among Ph.D. students. Sci. Rep. **11**(1), 14370 (2021). https://doi.org/10.1038/s41598-021-93687-7
23. Scheffel, M., Drachsler, H., Toisoul, C., Ternier, S., Specht, M.: The proof of the pudding: examining validity and reliability of the evaluation framework for learning analytics. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) EC-TEL 2017. LNCS, vol. 10474, pp. 194–208. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66610-5_15
24. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat Methodol. **58**(1), 267–288 (1996)
25. Wang, F., Hannafin, M.J.: Design-based research and technology-enhanced learning environments. Educ. Tech. Res. Dev. **53**(4), 5–23 (2005). https://doi.org/10.1007/BF02504682
26. Wollast, R., et al.: Who are the doctoral students who drop out? Factors associated with the rate of doctoral degree completion in universities. Int. J. High. Educ. **7**(4), 143 (2018). https://doi.org/10.5430/ijhe.v7n4p143

AGen: Personalized Analogy Generation with Large Language Model

Shutong Wu(✉), Hecong Wang, and Zhen Bai

University of Rochester, Rochester, NY 14620, USA
{swu85,hwang99}@ur.rochester.edu, zhen.bai@rochester.edu

Abstract. Analogies facilitate understanding across domains, enabling individuals to navigate unfamiliar concepts using what they already know. Yet, crafting effective analogies requires extensive knowledge of both the source and target domains, making it a challenging task for educators—particularly when aiming to support novice learners. Recent advancements in Large Language Models (LLMs) have demonstrated their ability to generate analogies to explain scientific concepts. However, analogy in education must align with students' prior knowledge and cognitive resources, which shape how analogies are perceived. To address this challenge, we examine the theory and practices of using analogies in education and introduce a system that generates personalized analogies to facilitate learning. Preliminary evaluations show promising outcomes, and we anticipate deeper insights from future human-subject studies.

Keywords: Analogical Learning · Personalization · Large Language Model

1 Introduction

Analogies are central and ubiquitous in human cognition, as people often explore new situations by drawing parallels to familiar ones [12]. In the educational domain, analogies are powerful tools for facilitating students' understanding of unfamiliar concepts and increasing motivation by introducing a familiar source domain [15]. Incorporating analogies into science texts and teaching has enhanced student comprehension, particularly by elucidating the causal and relational structures of concepts [19].

However, manually creating effective analogies is a cognitively demanding task, especially for novices. Research indicates that generating high-quality analogies requires effort, expertise, and a nuanced understanding of the subject matters of both the source and target domains [14,25]. Moreover, the effectiveness of analogies in learning is in line with the "Expertise Reversal Effect" framework [21], which highlights that with growing expertise and cognitive capacity, learners require different types of analogies to support their understanding best. For example, novice learners often benefit from surface-similar analogies, while

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 242–249, 2025.
https://doi.org/10.1007/978-3-031-99264-3_30

experts prefer structurally similar analogies that highlight deeper causal relationships [12,14].

Recent advancements in Generative AI and Large Language Models (LLMs) [24] have enabled the automatic generation of natural language analogies in settings such as education and creative writing [3,7,22]. These systems produce novel and meaningful analogies, but there is scarce research that takes into account the different ways audiences with varying backgrounds perceive and process analogies. One exception is the recent work aimed at generating personalized analogies for AI concepts [4]. While this work demonstrates the potential of tailoring analogies to user profiles, it primarily focuses on high-level prompting strategies and lacks a structured pipeline for extracting and adapting conceptual knowledge.

To address these challenges, we propose a novel application for generating personalized analogies based on users' prior experience and information such as learner age and personal interests. Our initial evaluation shows promising results and demonstrates the potential to improve analogy generation and learning outcomes. Future work includes comprehensive user studies to assess generation quality and practical usability.

2 Related Work

2.1 Analogy and Analogical Learning

Analogy is a fundamental cognitive tool that allows humans to learn across domains by leveraging prior knowledge [18]. Its significance has been recognized since ancient times, tracing back to Aristotle's theoretical reflections on analogy [1], and continues to influence modern understanding. Analogies are effective in various contexts, including science education [6,13], children's learning [16,23], and creativity-focused activities [9,20]. For novices, analogies are critical in enhancing problem-solving capabilities by facilitating solution transfer [1] and reducing uncertainty during problem-solving tasks [5].

Despite its potential, creating high-quality analogies often demands expertise [26], making it difficult for novices to perform meaningful inferences in unfamiliar domains without expert assistance [15]. Moreover, the effectiveness of an analogy can vary significantly across different age groups and audiences. For instance, young children are more likely to rely on object-level matches and surface similarities, whereas adults tend to focus on relational similarities between domains [10]. This *relational shift* underscores the importance of tailoring analogies to the target audience's cognitive capabilities and prior knowledge. Similarly, individuals with varied cognitive abilities [17] and relational knowledge [8] are likely to interpret and accept analogies differently based on their prior knowledge and cognitive resources.

2.2 Analogy Generation with Large Language Model

LLMs, such as GPT [24], have opened up many possibilities in various domains, including analogy generation [3] and analogical reasoning [27]. Previous studies

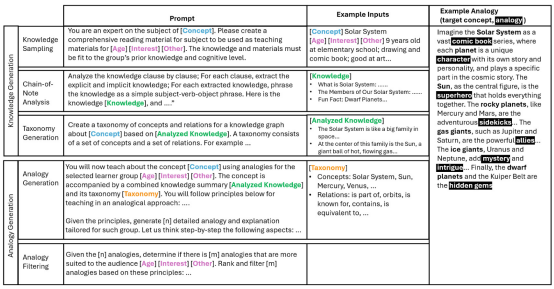


Fig. 1. Prompt Design Overview.

have demonstrated that LLMs can be effectively prompted to generate meaningful analogies and provide corresponding explanations when given specific contexts [3,9,22]. Researchers have extended LLM-based analogy generation beyond proportional analogies(A::B to C::D) by incorporating frameworks like Structure Mapping Theory [11] to generate high-level analogies [29]. While recent work has begun exploring personalization in analogy generation [4], it focuses primarily on modifying prompt templates for specific audiences and lacks a structured mechanism for representing domain knowledge or filtering analogies for educational applicability.

3 System Implementation

To address the limitations of existing work and achieve a more comprehensive and structured analogy generation, we introduce AGen, an analogy-generation tool that, given a source concept and an audience profile, creates tailored, structured analogies. We investigated previously established frameworks and principles in analogical learning [12,15], and incorporated these into our system to generate more structured and comprehensive analogies. Our framework consists of two stages, which are *Knowledge Generation and Adaptation* and *Analogy Generation and Selection*. We currently use GPT-4o as the underlying LLM, given its high performance on various tasks. Our step-by-step prompt design can be found in Fig. 1.

3.1 Knowledge Generation and Adaptation

To achieve effective personalization, we begin by collecting the audience profile across three dimensions: *learner age*, *background knowledge*, and *learner interest*, which is similar to the prompting framework in [4]. However, our system addresses a key limitation of prior approaches—the lack of distinction between conceptual relevance and motivational appeal—by treating these dimensions differently to solve distinct personalization challenges and achieve structural generation. Specifically, we use *learner age* and *background knowledge* as conditioning

inputs during the knowledge generation and adaptation stage. These factors shape a learner's cognitive readiness, domain familiarity, and ability to interpret relational structures—making them critical for generating analogies that are conceptually appropriate and structurally sound. In contrast, we use *learner interest* during the analogy generation and selection stage to filter and prioritize analogies based on motivational relevance. While interests can enhance engagement and retention, they should not interfere with the analogy's conceptual integrity.

Specifically, we implement three novel components—**Knowledge Sampling**, **Chain-of-Note Analysis**, and **Taxonomy Generation**—to shape analogy generation based on what the learner is likely to know and understand. We first sample the knowledge to develop a comprehensive knowledge of the source concept. We prompt GPT to generate background knowledge of a given science concept N times ($\mathrm{N} = 10$) and sample them to form a combined background information text. We then analyze the background knowledge to extract explicit and implicit knowledge as simple subject-verb-object phrases using the Chain-of-Note technique [28]. The analyzed knowledge is then aggregated into a taxonomy of the given concept, which consists of a set of filtered concepts and relations; in our early experiments, we found that a taxonomy is helpful in preventing LLMs from generating redundant or irrelevant information that could result in noise in the subsequent generation stages. We then adapt the generated taxonomy to the audience profile to generate a conditioned knowledge base that reflects the audience's prior knowledge of the source concept.

3.2 Analogy Generation and Selection

To generate the personalized analogy, we first prompt LLM to generate K analogies ($\mathrm{K} = 10$) based on the source concept, audience profile, and taxonomy. The LLM is prompted to create the analogy following the principles of analogical teaching [15], which includes principles such as using well-understood analogies and highlighting shared causal structure among analogies [15].

The above generation will still face limitations such as misinterpretations of certain sub-concepts or factually incorrect analogies. We further design the system to automatically filter the analogies by embedding the evaluation metrics in [2] and [12] into the prompts. Using the previously obtained taxonomy and additional user input, including *learner interest*, AGen filters the analogies only to output those easily understandable to the selected audience. An example of two tailored analogies towards different audiences is shown in Fig. 2.

4 Evaluation

We conducted a preliminary technical evaluation to test the personalization quality of the generated analogies in addressing intended audience groups. In the future, we will evaluate this tool's learning effectiveness and generation quality.

To test our tool, we selected a set of STEM science concepts from **Next Generation Science Standards (NGSS)**[1] by randomly sampling a group

[1] https://www.nextgenscience.org.

Example Analogy (target concept, analogy)		
6 years old, likes toys and cartoons like Peppa Pig	9 years old, likes drawing and comic book, good at art	16 years old, likes science and physics
Imagine the Solar System as a big, colorful carousel at the fair. The Sun is the tall pole in the middle, holding everything together, and each planet is like a fun horse that goes around and around. Mercury is the horse closet to the pole, and Neptune is way out on the edge. Each horse has a different color and style, just like each planet is special – Mars is red, Saturn has rings, and Uranus spins a little sideways! All the planets keep spinning around the Sun in a big circle, just like carousel horses moving together in a magical ride.	Imagine the Solar System as a vast comic book series, where each planet is a unique character with its own story and personality, and plays a specific part in the cosmic story. The Sun, as the central figure, is the superhero that holds everything together. The rocky planets, like Mercury and Mars, are the adventurous sidekicks... The gas giants, such as Jupiter and Saturn, are the powerful allies... The ice giants, Uranus and Neptune, add mystery and intrigue... Finally, the dwarf planets and the Kuiper Belt are the hidden gems	Imagine the Solar System as a giant atom. At its center, the Sun acts like the nucleus - a massive body that holds the system together.... Just as the nucleus is the dense core of an atom, the sun is the central powerhouse of the Solar System. Surrounding the Sun are the planets, akin to electrons orbiting the nucleus. These planets follow specific paths just like orbits, maintaining a delicate balance between the forces pulling them inward and those propelling them outward.

Fig. 2. AGen's Personalization Analogy Examples.

of concepts from different fields. This ensures that our experiments reflect the education curriculum. For further experiments, we will include more concepts.

4.1 Personalization Evaluation

We evaluate the system's ability to tailor analogies to audiences with different backgrounds and levels of expertise. We evaluated the system's personalization ability by classifying each analogy to the most likely audience profile. Specifically, we computed cosine similarities between the embeddings of each generated analogy and the embeddings of various audience profiles using the *text-embedding-3-small* model. The audience profile with the highest similarity score was selected as the predicted profile for each analogy. This approach allowed us to quantitatively assess how well the system aligns analogies with the intended audience characteristics.

The personalized evaluation examines the system's ability to tailor analogies based on audience profiles, specifically their ages and interests. Figure 3 provides confusion matrices summarizing system performance across these dimensions. Key findings are as follows: (1) **Strong alignment with user interests**: As shown in the right panel of Fig. 3, the system demonstrates high accuracy in predicting analogies aligned with users' declared interests. For instance, analogies tailored for individuals interested in comic books or sports are predominantly categorized correctly, with 14 accurate predictions in these categories. This strong alignment enhances the analogies' relevance, engagement, and potential impact on comprehension. However, minor misclassifications, such as sports analogies attributed to the "Animals" category, highlight areas for further refinement. (2) **Adherence to an audience age "upper bound"**: The left panel of Fig. 3 reveals that analogies respect an implicit "upper bound" for complexity based on the audience's age. Preschool-level analogies are correctly simplified

and context-specific, with the system achieving 12 accurate classifications for this group. Conversely, analogies for College Undergraduates exhibit greater breadth and complexity. Interestingly, while analogies designed for older audiences could, in principle, be simplified and adapted for younger audiences, the reverse does not hold, reflecting the system's capacity to generate age-appropriate outputs.

The evaluation highlights the system's strengths in aligning analogies with users' interests and age-appropriate complexity, as well as areas where further adjustments to mitigate biases can enhance performance.

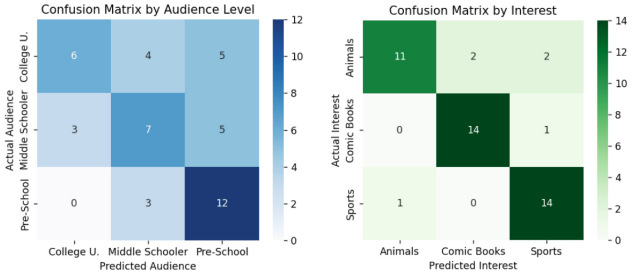


Fig. 3. Confusion Matrix based on two different settings.

5 Conclusion and Future Work

While AGen demonstrates strong potential, it also faces some limitations as an early-stage work. The reliance on LLMs for generating knowledge and taxonomy can result in incomplete or biased outputs due to inherent constraints in the model's pre-trained data, despite our efforts to condition the knowledge prospectively during the generation step. Moreover, evaluating the quality of analogies is inherently subjective, as it depends on users' prior knowledge and individual backgrounds. To address these challenges, future work will expand the scope of user studies and source concept domains to incorporate a more diverse group of participants, such as K-12 students and education experts.

This paper introduces AGen, an innovative analogy generation system that leverages advanced LLMs to create tailored, structured analogies for diverse educational contexts. By integrating knowledge adaptation and theory-driven generation methods, AGen represents a significant advancement in utilizing analogies to enhance learning.

Preliminary findings from our automatic evaluations show that our tool can produce personalized analogies with quality comparable to the existing baseline systems, indicating promising applications in improving concept comprehension and fostering user engagement. Future work will focus on conducting extensive user studies across varied demographics, exploring alternative and objective evaluation metrics, and integrating additional LLMs to enhance the diversity and

robustness of analogy generation. These efforts aim further to establish AGen as a valuable tool for educational innovation.

Acknowledgments. Thanks to Xiaofei Zhou for significantly assisting the research.

References

1. Bartha, P.: Analogy and Analogical Reasoning. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Fall 2024 edn. (2024)
2. Bhavya, B., Palaguachi, C., Zhou, Y., Bhat, S., Zhai, C.: Long-form analogy evaluation challenge. In: Mille, S., Clinciu, M.A. (eds.) Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, pp. 1–16. Association for Computational Linguistics, Tokyo (2024). https://aclanthology.org/2024.inlg-genchal.1/
3. Bhavya, B., Xiong, J., Zhai, C.: Analogy generation by prompting large language models: a case study of instructgpt (2022). https://arxiv.org/abs/2210.04186
4. CAO, C.C., Zoe, F., Lydia, Y.C., Jionghao, L., Ruizhe, L.: Llm-generated personalized analogies to foster ai literacy in adult novices. In: International Conference on Computers in Education (2024)
5. Chan, J., Paletz, S.B.F., Schunn, C.D.: Analogy as a strategy for supporting complex problem solving under uncertainty. Mem. Cogn. **40**(8), 1352–1365 (2012). https://doi.org/10.3758/s13421-012-0227-z
6. Coll, R.K., France, B., and, I.T.: The role of models/and analogies in science education: implications from research. Int. J. Sci. Educ. **27**(2), 183–198 (2005). https://doi.org/10.1080/0950069042000276712
7. Czinczoll, T., Yannakoudakis, H., Mishra, P., Shutova, E.: Scientific and creative analogies in pretrained language models (2022). https://arxiv.org/abs/2211.15268
8. D Gentner, M.R.: Language and the career of similarity. Technical Report 533, University of Illinois (1991)
9. Ding, Z., Srinivasan, A., Macneil, S., Chan, J.: Fluid transformers and creative analogies: exploring large language models' capacity for augmenting cross-domain analogical creativity. In: Proceedings of the 15th Conference on Creativity and Cognition, C&C '23, pp. 489–505. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3591196.3593516
10. Gentner, D.: Metaphor as structure mapping: the relational shift. Child Dev. **59**(1), 47–59 (1988). http://www.jstor.org/stable/1130388
11. Gentner, D., Markman, A.B.: Structure mapping in analogy and similarity. Am. Psychol. **52**(1), 45 (1997)
12. Gentner, D., Smith, L.A.: Analogical learning and reasoning. In: The Oxford Handbook of Cognitive Psychology, pp. 668–681. Oxford library of psychology, Oxford University Press, New York (2013). https://doi.org/10.1093/oxfordhb/9780195376746.001.0001
13. Gick, M., Holyoak, K.: Analogical problem solving. Cogn. Psychol. **12**, 306–355 (1980). https://doi.org/10.1016/0010-0285(80)90013-4
14. Goldwater, M.B., Gentner, D., LaDue, N.D., Libarkin, J.C.: Analogy generation in science experts and novices. Cogn. Sci. **45**(9), e13036 (2021)

15. Gray, M.E., Holyoak, K.J.: Teaching by analogy: from theory to practice. Mind Brain Educ. **15**(3), 250–263 (2021). https://doi.org/10.1111/mbe.12288. https://onlinelibrary.wiley.com/doi/abs/10.1111/mbe.12288
16. Haglund, J., Jeppsson, F., Andersson, J.: Young children's analogical reasoning in science domains. Sci. Educ. **96**(4), 725–756 (2012). https://doi.org/10.1002/sce.21009. https://onlinelibrary.wiley.com/doi/abs/10.1002/sce.21009
17. Halford, G.S.: Children's Understanding: The Development of Mental Models. Psychology Press (2014)
18. Hofstadter, D.R.: Epilogue: analogy as the core of cognition. In: The Analogical Mind: Perspectives from Cognitive Science. The MIT Press (2001). https://doi.org/10.7551/mitpress/1251.003.0020
19. Jaeger, A.J., Taylor, A.R., Wiley, J.: When, and for whom, analogies help: the role of spatial skills and interleaved presentation. J. Educ. Psychol. **108**(8), 1121 (2016)
20. Jiayang, C., et al.: Storyanalogy: deriving story-level analogies from large language models to unlock analogical understanding (2023). https://arxiv.org/abs/2310.12874
21. Kalyuga, S.: Expertise reversal effect and its implications for learner-tailored instruction. Educ. Psychol. Rev. **19**, 509–539 (2007)
22. Kim, J., Suh, S., Chilton, L.B., Xia, H.: Metaphorian: leveraging large language models to support extended metaphor creation for science writing. In: Proceedings of the 2023 ACM Designing Interactive Systems Conference, DIS '23, pp. 115–135. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3563657.3595996
23. Newton, D.P., Newton, L.D.: Using analogy to help young children understand. Educ. Stud. **21**(3), 379–393 (1995). https://doi.org/10.1080/0305569950210305
24. Ouyang, L., et al.: Training language models to follow instructions with human feedback (2022). https://arxiv.org/abs/2203.02155
25. Richland, L.E., Simms, N.: Analogy, higher order thinking, and education. Wiley Interdisc. Rev. Cogn. Sci. **6**(2), 177–192 (2015)
26. Turney, P.D.: The latent relation mapping engine: algorithm and experiments. J. Artif. Intell. Res. **33**, 615–655 (2008). https://doi.org/10.1613/jair.2693
27. Yasunaga, M., et al.: Large language models as analogical reasoners (2024). https://arxiv.org/abs/2310.01714
28. Yu, W., Zhang, H., Pan, X., Ma, K., Wang, H., Yu, D.: Chain-of-note: enhancing robustness in retrieval-augmented language models (2024). https://arxiv.org/abs/2311.09210
29. Yuan, S., Chen, J., Ge, X., Xiao, Y., Yang, D.: Beneath surface similarity: large language models make reasonable scientific analogies after structure abduction (2023). https://arxiv.org/abs/2305.12660



'The World of AI': A Novel Approach to AI Literacy for First-Year Engineering Students

Siddharth Siddharth(✉), Brainerd Prince, Amol Harsh, and Shreyas Ramachandran

Plaksha University, Mohali, India
{siddharth.s,brainerd.prince,amol.harsh,
shreyas.ramachandran}@plaksha.edu.in

Abstract. This work presents a novel course titled 'The World of AI' designed for first-year undergraduate engineering students with little to no prior exposure to AI. The central problem addressed by this course is that engineering students often lack foundational knowledge of AI and its broader societal implications at the outset of their academic journeys. We believe the way to address this gap is to design and deliver an interdisciplinary course that can a) be accessed by first-year undergraduate engineering students across any domain, b) enable them to understand the basic workings of AI systems sans mathematics, and c) make them appreciate AI's far-reaching implications on our lives. The course was divided into three modules co-delivered by faculty from both engineering and humanities. The planetary module explored AI's dual role as both a catalyst for sustainability and a contributor to environmental challenges. The societal impact module focused on AI biases and concerns around privacy and fairness. Lastly, the workplace module highlighted AI-driven job displacement, emphasizing the importance of adaptation. The novelty of this course lies in its interdisciplinary curriculum design and pedagogical approach, which combines technical instruction with societal discourse. Results revealed that students' comprehension of AI challenges improved across diverse metrics like (a) increased awareness of AI's environmental impact, and (b) efficient corrective solutions for AI fairness. Furthermore, it also indicated the evolution in students' perception of AI's transformative impact on our lives.

Keywords: education about AI · teaching AI · undergraduate engineering education · AI literacy · critical reflection on AI applications

1 Introduction

Artificial Intelligence (AI) is reshaping technology, society, and the workplace [15], yet many first-year engineering students lack foundational knowledge of AI's societal impact and ethical challenges, highlighting the need for comprehensive

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 250–257, 2025.
https://doi.org/10.1007/978-3-031-99264-3_31

ethics training [3,5,8]. In response, the interdisciplinary 'World of AI' course was developed for first-year students; unlike traditional AI courses that emphasize technical and mathematical rigor, this course demystifies AI and focuses on its social implications, as research indicates that combining technical instruction with humanities perspectives enhances both AI proficiency [12] and ethical understanding [18]. Thus, the 'World of AI' course uniquely integrates STEM and humanities to equip students with essential technical skills and a critical awareness of AI's broader societal impacts. The course comprises of three modules—AI and the Planet, AI and the Society, and AI and the Workplace—and introduces foundational AI concepts and ethics, emphasizing its dual nature as both a driver of innovation and a source of challenges, such as environmental stress [9], societal biases [10], and workplace disruptions, thereby prompting a critical evaluation of its opportunities and limitations. Collaboratively taught by STEM and humanities faculty, it addresses overlooked themes in traditional curricula and is evaluated via pre-course and post-course surveys and case studies examining AI's implications on the planet, society, and the workplace.

2 Background

Many first-year engineering students often lack foundational knowledge of AI and its broader societal implications when they begin their academic journeys. This gap in understanding has been documented across numerous studies, which highlight how traditional engineering curricula prioritize technical knowledge at the expense of societal and ethical perspectives [2]. Research has equally shown the negative impacts of developing and using AI technology. Environmental implications of AI have drawn increasing attention in recent years. Studies by Strubell [14] emphasize the significant energy consumption associated with training large AI models, with data centers contributing substantially to global carbon emissions. These findings underscore the need for future engineers to recognize the environmental trade-offs of AI technologies and to innovate solutions that mitigate such impacts. However, students are neither taught nor exposed to these ecological costs of AI, thus leaving a crucial gap in their education. In the social domain, research by Obermeyer [7] has illustrated how biases inherent in AI systems can perpetuate inequalities, particularly in sensitive areas like healthcare, criminal justice, and employment. This raises concerns about fairness and equity, which are often overlooked in traditional engineering curricula [6]. The work on the impact of AI on employment by Frey and Osborne [4] argue that advancements in AI-driven automation have the potential to displace a significant portion of the workplace, particularly in industries reliant on routine tasks. At the same time, AI creates opportunities for new job roles, demanding a workplace that is adaptable and equipped with interdisciplinary skills. Despite these drastic recent shifts, engineering programs have not yet addressed how students can prepare for such disruptions, leaving them unprepared for the changing dynamics of the workplace [1].

In response to this shortfall, several universities and online platforms now offer stand-alone subjects on AI ethics or literacy, confirming rising

demand yet leaving first-year engineers underserved. Coursera's open MOOC *Ethics of Artificial Intelligence*, delivered by Politecnico di Milano, reaches a global audience but provides no project component and is not embedded in any engineering curriculum [11]. Stanford's flagship CS/ETHICSOC 182 blends philosophy, public policy, and computer science but requires CS 106A, attracting mainly second- or third-year CS students [17]. By contrast, our *World of AI* is a required gateway in semester 1, co-taught by humanities and engineering faculty, and uses case study and projects to couple foundational concepts with environmental, social, and workplace analysis—an approach we have not found in any other entry-level engineering programme.

3 Methodology

Research Question. The central research question addressed in this paper is: to what extent does this introductory course on AI impact the first-year engineering undergraduate students' understanding about the role and impact of AI on the planet, the society, and the workplace?

Course Design. The course blended technical instruction with industry insights and interdisciplinary perspectives, guiding students from basic AI concepts to a critical evaluation of its societal implications through a series of progressive modules. Initially, three engineering lectures demystified key concepts like machine learning and neural networks, laying the groundwork for subsequent discussions, which were enriched by guest lectures from industry leaders and co-delivered sessions by engineering and humanities faculty.

The curriculum was organized into three main modules:

- **AI and the Planet** This module examined AI's dual role in driving sustainable progress while posing environmental challenges through high energy demands and extensive water use for data center cooling, and also highlighted AI's potential to optimize renewable resource management.
- **AI and the Society** This module explored societal biases in decision-making, risks of biased AI model training, and the balance between personalization and privacy while emphasizing fairness, equity, and inclusivity.
- **AI and the Workplace** This module analyzed the evolving dynamics in an AI-driven economy, addressing both the emergence of new industries and the displacement of traditional roles, and underscored the need for reskilling initiatives to help workers adapt to technological change.

The course was delivered in person through two weekly 1-hour classroom sessions, supported by regular office hours with instructors and dedicated TA-led sessions for resolving doubts. This structure facilitated active engagement, consistent feedbacks.

Data Collection. This study employed a mixed-method, single-group pre-post design that combined quantitative multiple-choice surveys with qualitative open-ended responses. A total of 168 first-year engineering undergraduates (113 male, 55 female; ages 17–19) participated in a 14-week in-person elective course, with all participants providing informed consent for anonymized survey responses. Pre-course surveys, which included both multiple-choice and open-ended questions, established a baseline of students' understanding of AI's environmental, social, and workplace challenges and their ethical reasoning, probing their perceptions of AI's contemporary impact. At the end of the course, a post-course survey was administered via Microsoft Forms during the final lecture, allowing students 15–20 min to reflect on their learning and to capture shifts in their knowledge and ethical perspectives after completing the modules. To support construct validity, all pre- and post-course questionnaire items were adapted from established instruments, including the Meta AI Literacy Scale (MAILS) [13] and UNESCO's AI Ethics Readiness Assessment [16]. Ethical approval for the data collection was obtained from a research ethics committee.

In addition to the surveys, a rubric-based evaluation of the final group presentations measured learning outcomes; in these presentations, each group envisioned a future problem (by 2030) impacting the planet, society, or workplace and proposed an AI-based solution. The presentations were assessed on five criteria—Understanding of Future Problem (clarity in explaining the problem, its causes, and impact), Future Relevance (significance over the next five years), AI's Contribution to the Problem (how AI might exacerbate or influence the issue), Corrective AI Solution (feasibility of using AI to address the problem), and Responsibility & Safeguards (ethical risks and safety measures)—with each parameter rated on a 0–10 scale to enable performance comparisons across groups.

Preliminary Data Analysis. The quantitative scoring methodology provided standardized measures across question types: for objective questions, correct responses earned 1 point and incorrect responses earned 0; in MCQs, each correctly selected option received 1 point while unselected or incorrect options scored 0; and scoring in Likert-scale responses was "Strongly Agree" = 5, "Agree" = 4, "Neutral" = 3, "Disagree" = 2, and "Strongly Disagree" = 1, with raw scores subsequently scaled to a 10-point scale. Qualitative responses were evaluated using an automated LLM-based method that compared each answer to an ideal response, normalizing similarity scores to a 10-point scale, with a sample of responses spot-checked by evaluators. Final presentations were assessed by an external evaluator using a five-category rubric, where each criterion was rated on a scale from 1 (poor) to 10 (excellent) and weighted equally (Fig. 1).

4 Results

Quantitative assessment of course effectiveness through pre-course survey vs. final presentation scores: In order to systematically measure learning outcomes, we organized the pre-survey questions into five categories corresponding to the

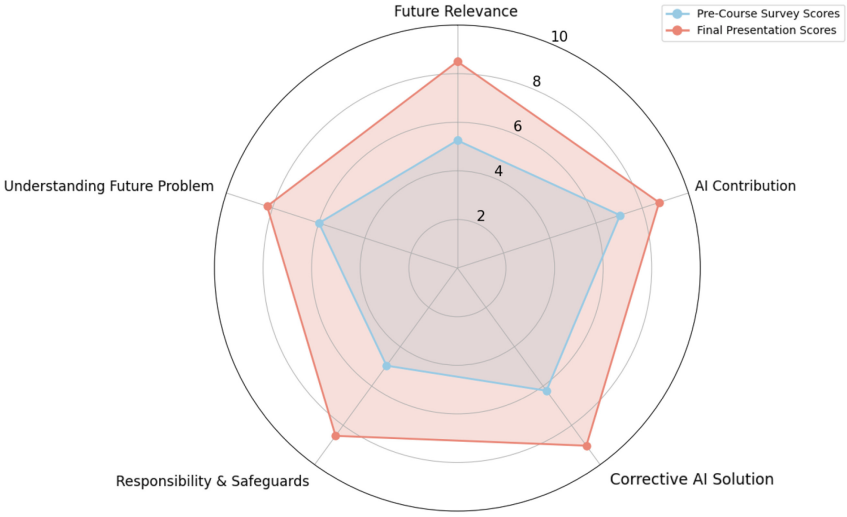


Fig. 1. Radar Chart: Pre-Course Survey vs. Final Presentation Scores.

final project rubric: Understanding the Future Problem, Future Relevance, AI's Contribution to the Problem, Corrective AI Solution, and Responsibility & Safeguards. These categories aligned with the rubrics outlined in the Methodology section, allowing for direct comparisons between students' initial responses and their final presentation scores. Specifically, the "Understanding the Future Problem" category was linked to questions on reducing AI's environmental impact and ensuring unbiased service. "Future Relevance" focused on balancing AI's benefits with sustainability by 2030, while "AI's Contribution to the Problem" addressed energy requirements for data centers and the importance of prioritizing environmental impacts. "Corrective AI Solution" involved modifying an AI tool for monitoring pollution in diverse regions, and "Responsibility & Safeguards" covered workplace bias and equitable AI-driven community recommendations.

In assessing various rubrics, students showed notable improvements from pre-survey to final presentation evaluations. By mapping these questions to the final rubric scores, we obtained a clear view of students' progress across all five domains. For **Understanding the Future Problem (Rubric 1)**, an open-ended question addressing real-world AI challenges such as energy consumption and algorithmic bias, the mean score increased from 5.3 (SD = 3.0) to 8.2 (SD = 1.1). In **Future Relevance (Rubric 2)**, where students evaluated a future scenario emphasizing the urgency of sustainable AI within the next five years, scores improved from 7.0 (SD = 2.3) to 8.5 (SD = 1.5). **AI's Contribution to the Problem (Rubric 3)** was gauged using two multiple-choice questions—one regarding the annual energy usage of AI-supporting data centers (correct answer: "Similar to global aviation industry energy consumption") and another on the environmental impact of these centers—with normalized scores rising from 6.2

($SD = 2.6$) to 8.7 ($SD = 0.8$). For **Corrective AI Solution (Rubric 4)**, which involved proposing enhancements to the AI tool's coverage while ensuring fair data representation, the average score went from 5.0 ($SD = 3.1$) to 9.0 ($SD = 0.9$). Finally, **Responsibility and Safeguards (Rubric 5)**—assessed through two multiple-choice questions addressing potential biases in workplace areas (hiring, performance evaluation, employee monitoring, task automation) and equitable AI recommendations in urban planning (with the correct response being to expand data collection in underrepresented communities)—saw scores increase from 5.9 ($SD = 2.7$) to 8.5 ($SD = 1.2$), reflecting a strong overall gain in students' awareness of fairness and ethical safeguards.

Because the gain scores were non-normal (Shapiro–Wilk, $p < .05$ for every rubric), we used paired Wilcoxon signed-rank tests (W = test statistic, $n = 168$) and applied a Bonferroni correction to control the family-wise error (p_{adj}). All five learning outcomes improved significantly: *AI Contribution* ($W = 1288.0$, $p_{\text{adj}} = 6.03 \times 10^{-13}$, $r = .47$), *Responsibility & Safeguards* ($W = 1012.0$, $p_{\text{adj}} = 1.21 \times 10^{-17}$, $r = .68$), *Corrective AI Solution* ($W = 471.0$, $p_{\text{adj}} = 1.16 \times 10^{-22}$, $r = .77$), *Future Relevance* ($W = 1717.0$, $p_{\text{adj}} = 5.30 \times 10^{-14}$, $r = .55$), and *Understanding Future Problem* ($W = 954.0$, $p_{\text{adj}} = 7.57 \times 10^{-19}$, $r = .68$). Here, p_{adj} is the Bonferroni-adjusted two-tailed probability, and r is the rank-biserial effect size ($|r| \geq .50$ = large). Four rubrics show large practical gains ($r = .55$–$.77$) and one a moderate–large gain ($r = .47$), confirming that the observed improvements are both statistically robust and educationally meaningful.

Change in learners perspective from post-course survey: The post-survey, illustrated in Fig. 2, captured detailed shifts in students' understanding of AI across key areas—from technical distinctions between AI and traditional programming to its environmental, societal, and workplace impacts—through 7 Likert-scale questions that quantified changes since the course began. Questions addressed topics such as **AI & the Planet**, asking to rate the effectiveness of AI-optimized energy solutions (e.g., nuclear for reliability and low-carbon output, solar for renewability despite intermittency) in balancing energy demands with sustainability; **AI & the Society**, evaluating the ability of AI to balance personalized experiences with safeguarding user privacy; and **AI & the Workplace**, assessing how effectively the course helped overcome fears of job displacement. The highest mean rating, approximately 7.5, was for understanding the difference between AI and traditional programming, while balancing personalization with AI systems' adaptability received a mean score around 6.5, indicating moderate confidence. Overall, the responses reflect a positive shift in students' perceptions across multiple AI domains.

5 Discussion

The central question of this study—how an introductory AI course impacts first-year engineering students' understanding of AI's role in the planet, society, and the workplace—is affirmatively answered through the course assessments in this research. Across all five rubrics, students demonstrated marked improvements

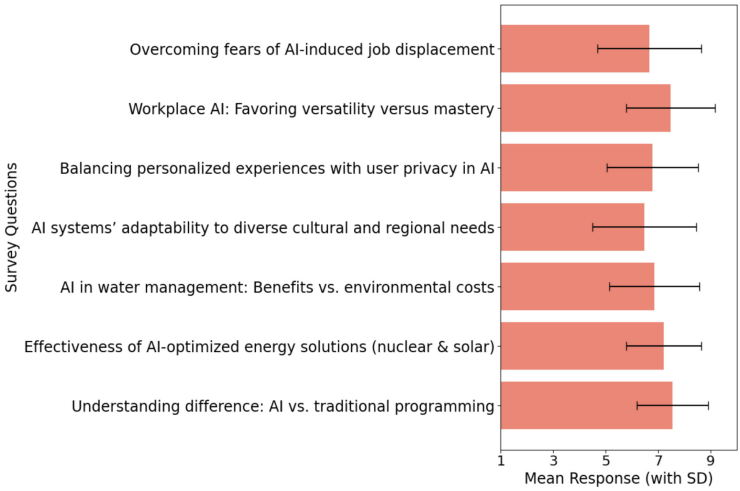


Fig. 2. Students' post-course survey responses showing the extent of change in their understanding of various AI aspects (1 = Not at all, 10 = Completely).

in final presentations compared to baseline scores, while the post-course survey revealed moderate positive shifts in perceptions ranging from technical distinctions (machine learning versus traditional programming) to broader issues such as AI's role in sustainability, with average ratings between approximately 6.5 and 7.5 on a 10-point scale (see Fig. 2). Nonetheless, several limitations warrant caution: the reliance on self-reported data introduces potential bias and subjective interpretation, and the absence of a control group or longitudinal follow-up leaves it unclear whether these attitudinal changes can be solely attributable to the course. Future work could explore longer-term impacts on students' professional trajectories by tracking the cohort through subsequent semesters, internships, and early career experiences to better assess the enduring influence of the course's conceptual foundations and ethical frameworks.

Acknowledgments. The authors are thankful to Harish and Bina Shah School of AI & CS at Plaksha University for providing the seed financial support for this research.

References

1. Autor, D.H.: Why are there still so many jobs? The history and future of workplace automation. J. Econ. Perspect. **29**(3), 3–30 (2015)
2. Baker, T., Smith, L.: Educ-ai-tion rebooted? exploring the future of artificial intelligence in schools and colleges. Technical report, NESTA (2019)
3. Feffer, M., Martelaro, N., Heidari, H.: The ai incident database as an educational tool to raise awareness of ai harms: a classroom exploration of efficacy, limitations, & future improvements. arXiv preprint arXiv:2310.06269 (2023)

4. Frey, C.B., Osborne, M.A.: The future of employment: how susceptible are jobs to computerisation? Technol. Forecast. Soc. Chang. **114**, 254–280 (2017)
5. Goldenkoff, E., Cech, E.A.: Left on their own: confronting absences of ai ethics training among engineering master's students. In: Proceedings of the 2024 ASEE Annual Conference & Exposition. American Society for Engineering Education (2024). https://doi.org/10.18260/1-2--47724
6. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021)
7. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342
8. Orchard, A., Radke, D.: An analysis of engineering students' responses to an ai ethics scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 13, pp. 15834–15842 (2024). https://doi.org/10.1609/aaai.v37i13.26880. https://ojs.aaai.org/index.php/AAAI/article/view/26880
9. Patterson, D., et al.: Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350 (2021)
10. Pillsbury, W.S.P.L.: The impact of artificial intelligence on vulnerable populations in the workforce. Pillsbury Law (2022). https://www.pillsburylaw.com/en/news-and-insights/ai-impact-vulnerable-workforce.html
11. Politecnico di Milano: Ethics of artificial intelligence (2024). https://www.coursera.org/learn/ethics-of-artificial-intelligence. Massive Open Online Course on Coursera. Accessed 5 May 2025
12. Stadelmann, T., Keuzenkamp, J., Grabner, H., Würsch, C.: The ai-atlas: didactics for teaching ai and machine learning on-site, online, and hybrid. Educ. Sci. **11**(7) (2021). https://doi.org/10.3390/educsci11070318. https://www.mdpi.com/2227-7102/11/7/318
13. Straka, S., Latoschik, M.E., Wienrich, C.: MAILS – meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Comput. Human Behav. Artif. Humans **1**(2), 100014 (2023). https://doi.org/10.1016/j.chbah.2023.100014
14. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in nlp. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650 (2019)
15. Tai, M.C.T.: The impact of artificial intelligence on human society and bioethics. Tzu Chi Med. J. **32**(4), 339–343 (2020). https://doi.org/10.4103/tcmj.tcmj_71_20
16. UNESCO: Readiness assessment methodology: A tool of the recommendation on the ethics of artificial intelligence. Technical report, United Nations Educational, Scientific and Cultural Organization, Paris (2023). https://www.unesco.org/en/articles/readiness-assessment-methodology-tool-recommendation-ethics-artificial-intelligence
17. University, S.: Cs/ethicsoc 182: Ethics, public policy, and technological change (2025). https://web.stanford.edu/class/cs182/. Course syllabus. Accessed 5 May 2025
18. Usher, M., Barak, M.: Integrating ai ethics into science and engineering curricula: a case for explicit-reflective learning. Int. J. STEM Educ. **11**(3), 45–60 (2024). https://doi.org/10.1186/s40594-024-00493-4

Exploring Communication Strategies for Collaborative LLM Agents in Mathematical Problem-Solving

Liang Zhang[1,2], Xiaoming Zhai[3,4,5(✉)], Jionghao Lin[6], Jennifer Kleiman[3,5], Diego Zapata-Rivera[7], Carol Forsyth[7], Yang Jiang[7], Xiangen Hu[8], and Arthur C. Graesser[9]

[1] Institute for Intelligent Systems, University of Memphis, Memphis, TN, USA
lzhang13@memphis.edu
[2] Department of Electrical and Computer Engineering, University of Memphis, Memphis, TN, USA
[3] AI4STEM Education Center, University of Georgia, Athens, GA, USA
Xiaoming.Zhai@uga.edu
[4] National GENIUS Center, Athens, GA, USA
[5] Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA
[6] Faculty of Education, The University of Hong Kong, Hong Kong, People's Republic of China
[7] Educational Testing Service, Princeton, NJ, USA
[8] Department of Applied Social Sciences, Hong Kong Polytechnic University, Hong Kong, People's Republic of China
[9] Department of Psychology, University of Memphis, Memphis, TN, USA

Abstract. Large Language Model (LLM) agents are increasingly used in AI-supported education. Effective inter-agent communication boosts collaborative problem-solving efficiency and lowers the cost of deploying LLM-driven educational systems. However, few studies have systematically examined how different communication strategies affect on agents' problem-solving performance, which is essential for developing effective collaborative applications with LLM agents. Inspired by principled scenarios in learning science and engineering, our study examines four communication modes, *teacher-student interaction*, *peer-to-peer collaboration*, *reciprocal peer teaching*, and *critical debate*, in a dual-agent, chat-based mathematical problem-solving environment powered by OpenAI's GPT-4o model. On the MATH benchmark, our results show that dual-agent setups outperform single agents, with *peer-to-peer collaboration* achieving the highest accuracy. Dialogue acts like statements, acknowledgment, and hints play a key role in collaborative problem-solving. These findings show that multi-agent frameworks can boost collaborative performance, but carefully crafted effective communication strategies remain essential for tackling complex tasks in AI-supported education.

Keywords: Large Language Model · Dual Agents · Communication Strategies · Math Problem Solving · Artificial Intelligence

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 258–265, 2025.
https://doi.org/10.1007/978-3-031-99264-3_32

1 Introduction

Recent advances in Large Language Models (LLMs) have shown significant promise in enhancing educational tasks such as feedback generation, learning guidance, interaction strategies, understanding student behaviors, and fostering tutoring dialogues through answer evaluation and content generation [13,20,22,23]. However, most implementations rely on single-agent approaches using prompt engineering for logical reasoning (e.g., Chain of Thought [17]). Emerging research on multi-agent collaboration among LLMs further improves performance through collaborative strategies and holds promise for expanding the effectiveness of AI applications in education [18,19]. In multi-agent settings, LLM agents leverage complementary reasoning processes to cross-validate ideas, refine their understanding, and ultimately produce more robust solutions than single-agent approaches [7]. Despite recent advances, systematic studies on how communication strategies shape the performance of LLM-powered multi-agent systems remains unexplored, particularly in settings where collaborating agents assume distinct personas and task-specific roles.

LLM-powered multi-agent systems can potentially mimic the key advantages of human collaboration by sharing and cross-validating diverse perspectives, thus boosting efficiency and sparking emergent reasoning [1,9]. To assess and refine these collaborative benefits, we need a domain with rigorous, verifiable metrics. Mathematics is an ideal testbed, as its step-wise logic and binary correctness offer rigorous, transparent metrics. Recent work on chat-based tutoring [10] and multi-agent conversation programming for math problem-solving [5] already shows coordinated LLMs tackling the field's unique reasoning challenges. A systematic study of collaborative strategies in this domain therefore promises actionable design principles, optimizing agent roles, turn-taking protocols, and feedback mechanisms, that can generalize to AI-driven educational systems [21].

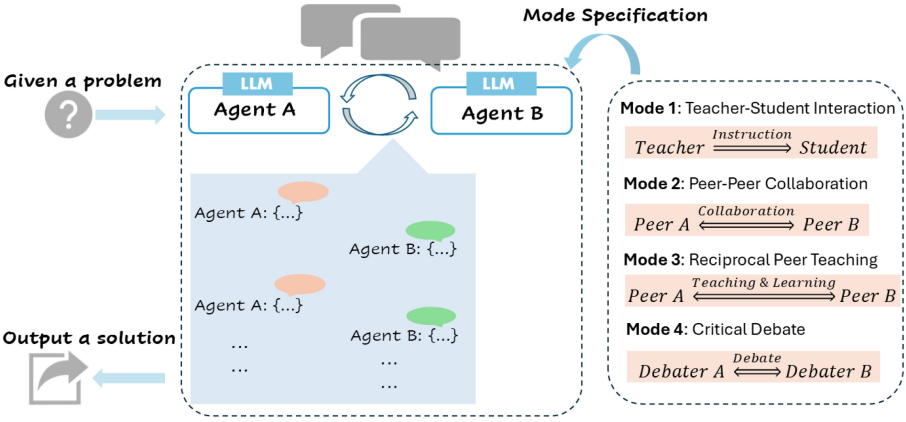


Fig. 1. An illustration for different dual LLM agents modes for problem-solving.

In this study, we explore a range of collaborative communication strategies using dual LLM-based agents to solve mathematical problems via chat-based interactions. Drawing inspirations from Vygotsky's educational theories, particularly the Zone of Proximal Development (ZPD) and the More Knowledgeable Other (MKO) [16], and informed by insights from machine psychology [2], we adopt several interaction modes in our study. These include *teacher-student interactions*, where one agent guides another through direct instruction and the Socratic method; *peer-to-peer collaboration*, in which agents work as equals to exchange ideas and solve problems; *reciprocal peer teaching*, where agents alternate roles of teacher and learner to reinforce understanding; and *critical debate*, where agents challenge one another's solutions to refine their approaches. Figure 1 illustrate further details on these communication modes. By assigning distinct personas to the LLM agents, we aim to enhance their collaborative problem-solving effectiveness and evaluate the performance of these communication strategies in math problem-solving. Dialogue Act (DA) analysis will be applied to conversations across different strategies to identify unique traits and derive actionable insights. Our investigation is guided by the following **R**esearch **Q**uestions:

- **RQ 1:** Which communication strategies among dual-agent LLM systems most effectively enhance mathematical problem-solving performance?
- **RQ 2:** How do DA patterns vary across dual-agent communication modes and influence mathematical problem-solving?

The study will inform the design of collaborative LLM agents in multi-agent settings, yielding generalized principles that enable more adaptive AI educational platforms. By identifying communication strategies that most enhance joint reasoning, the study guides the selection and orchestration of agents in intelligent tutoring systems. These insights bridge foundational multi-agent AI research and practical educational applications, ultimately improving learning outcomes in AI-supported education.

2 Method

All LLM agents are powered by the OpenAI GPT-4o model [4] and were evaluated on 700 level-5 (highest-difficulty) problems from MATH dataset[1] [3]. It includes seven different objects including algebra, counting and probability, geometry, intermediate algebra, number theory, prealgebra, and precalculus. Each setup was repeated in three independent runs to ensure reliability.

Dual-Agent Protocol Definition. We define four dual-agent communication protocols, each tailored to a distinct communication mode (see Fig. 1). Each protocol involves two LLM agents, $\mathcal{A}$ and $\mathcal{B}$, to whom we assign specific personas and provide comprehensive guidance for response generation via prompt engineering. For example, in the *teacher–student interaction* setup, the agent $\mathcal{A}$

[1] https://huggingface.co/datasets/hendrycks/competition_math.

adopts the teacher persona with the summarized version of prompt template $\mathcal{T}_A(Q, \mathcal{H}) =$ "*You are a supportive math teacher. Present Q clearly, summarize key points, offer hints, and ask probing questions based on dialogue history $\mathcal{H}$ without revealing the full solution*", and the agent $\mathcal{B}$ adopts the student persona with the summarized version of prompt template $\mathcal{T}_B(Q, \mathcal{H}) =$ "*You are a curious math student. Follow the teacher's guidance to solve Q step-by-step, articulate your reasoning, and ask for clarifications as needed.*" For further details, including comprehensive agent definitions and prompt templates for alternative communication modes, please refer to our GitHub repository:https://github.com/LiangZhang2017/aied_dual_agent.

Baseline. We established a baseline single-agent protocol in a zero-shot setting. The single agent is defined as a helpful assistant tasked with solving math problems, generating both the solution process and the final answer by sequentially working through each problem.

Accuracy Evaluation. Correctness is measured via exact-match comparison with ground truth. Accuracy, the proportion of correct answers on a 0–100 scale, is computed for each subject, communication mode, and run. We then report the mean across three independent runs and its standard error (SE) to capture variability.

Dialogue Act Classification Analysis. The conversation exhibits in-depth reasoning that reveals the underlying logical structure of mathematical problem solutions. To further examine the nuances of dialogue communication and explore the distinct functionalities of dual agents in collaborative math problem solving, we perform a Dialogue Act (DA) analysis. Dialogue acts refer to the functional roles that utterances or responses play in a conversation [11,12,14]. Dialogue acts can include actions like asking questions, providing explanations, and offering feedback. Initially, each agent's response is segmented into chunks, where a chunk may consist of a single sentence, a series of strongly logically related sentences, or a comprehensive multi-step equation process. We then used a pre-trained Bert-based model introduced in [11], which has been trained and validated on large-scale tutoring dialogue data (3,629 utterances) collected from math tutoring sessions between human tutors and students. We further applied a framework based on the principles from [14], which provides a structured method for categorizing different types of dialogue acts. The following is a brief reference list of DA tags from the scheme in [14]: H (Hint), DIR (Directive), ACK (Acknowledge), RC (Request Confirmation), RF (Request Feedback), PF (Positive Feedback), NF (Negative Feedback), LF (Lukewarm Feedback), Q (Question), A (Answer), and S (Statement).

3 Results and Discussion

3.1 Performance Comparison of Four Communication Modes

Table 1 presents the mean accuracy along with corresponding SE values for each communication mode averaged across all problem categories. All dual-agent

Table 1. Mean Accuracy (%) by Communication Mode (±SE)

Model	Single Agent	Dual Agents			
		Teacher-Student	Peer-to-Peer	Critical Debate	Reciprocal Peer
GPT-4o	$47.43_{6.59}$	$52.33_{7.43}$	$54.38_{7.09}$	$52.76_{7.44}$	$51.95_{7.12}$

configurations consistently outperform the single-agent baseline (47.43%), confirming that collaborative interactions enhance mathematical problem-solving accuracy. Notably, the *Peer-to-Peer collaboration* mode, characterized by agents equally sharing intermediate results and mutually verifying outputs, achieved the highest average accuracy (54.38%) and exhibited the lowest SE (7.09), indicating reliable performance. Matched pair-wise t-tests further confirmed that the *Peer-to-Peer collaboration* significantly outperformed the Single Agent mode ($t = 6.54, p = 0.0006$), highlighting the value of reciprocal and equal-role interactions in enhancing computational reasoning and error-checking essential for complex mathematical tasks [8]. The *Teacher-Student* ($t = 4.29, p = 0.0051$), *Critical Debate* ($t = 3.25, p = 0.0175$) and *Reciprocal Peer* ($t = 5.34, p = 0.0018$) modes also demonstrated statistically significant improvements over the single-agent baseline, although their effectiveness was less pronounced compared to *Peer-to-Peer* collaboration. Given the pairwise matched t-test results, the *Peer-to-Peer collaboration* mode is also significantly more effective than other dual-agent modes, notably surpassing the *Teacher-Student* mode ($t = 3.03$, $p = 0.0232$) and the *Reciprocal Peer* mode ($t = 3.20$, $p = 0.0186$).

3.2 Dialogue Act Analysis Result

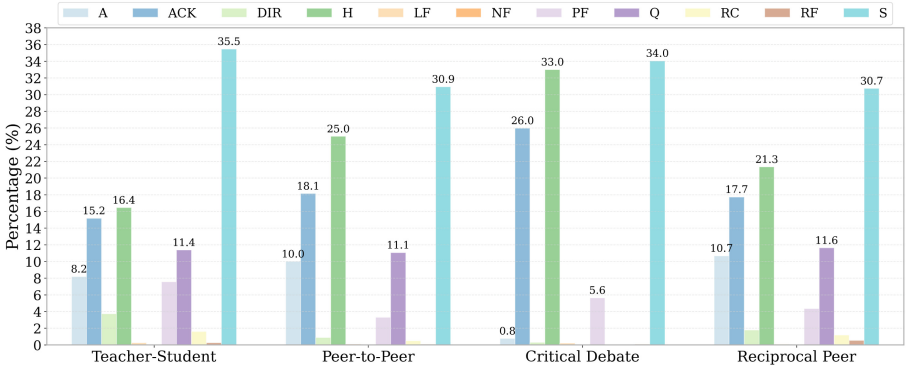


Fig. 2. DA Classification Results Across Four Modes.

Figure 2 illustrates the frequency distribution of DAs across four communication modes, exemplified by the algebra subject from MATH problems. Both agents'

utterances were merged to capture their collective impact on problem solving, and the top five DA tags in each mode are highlighted in the upper histogram. Notably, statement (S), hints (H) and acknowledge (ACK) emerge as the three most frequent dialogue acts across all modes, significantly contributing to collaborative problem-solving. In particular, statement (S) is the predominant DA across all four modes, with its peak frequency (35.5%) observed in the teacher-student interaction mode. For example, an agent might state, "*The algebraic manipulation and the verification steps are clear and logical*," or "*However, I would like to explore the solution further to ensure there are no other possible pairs of x and y that satisfy the given conditions.*" These examples illustrate how agents express their own perspectives and guide the subsequent moves in problem solving. The hints (H) emerge as the most frequent dialogue act in critical debate (33.0%), where agents actively challenge each other's solutions. The utterances "*Let's denote the two whole numbers I pick as x and y.*" and "*According to the problem, my friend picks the numbers x-4 and 2y-1.*" serve as hints that are intended to guide the problem-solving process rather than provide complete solutions. This high prevalence of hints reflects an active effort to explore, test, and improve upon proposed solutions [15]. Acknowledge (ACK), for instance, "*Thank you, Agent B, for your systematic approach to verifying the solution,*" or "*Thank you for the constructive discussion!*" serve as bridging elements that affirm successful communication between agents. These acknowledgments not only recognize individual contributions but also strengthen the overall collaborative dynamic. Additionally, the dialogue acts classified as answer (A) are among the top five most frequent categories. Overall, the high frequency of these dual-agent-specific dialogue acts underscores their crucial role in driving the collaborative problem-solving process. Using Kendall's correlation [6], we found that DA patterns in peer-to-peer collaboration are highly correlated with teacher-student interaction (0.99) and reciprocal peer teaching (0.96), but less so with critical debate (0.67), suggesting that even subtle variations in these patterns can significantly impact performance.

4 Conclusion

In this study, we investigated how different communication modes among LLM-based multi-agent systems affect mathematical problem-solving performance. By systematically comparing four distinct dual-agent communication modes including *teacher-student interaction*, *peer-to-peer collaboration*, *critical debate*, and *reciprocal peer teaching*, we demonstrated that multi-agent frameworks can significantly enhance problem-solving performance relative to single-agent approaches. Notably, the *peer-to-peer collaboration* mode achieved the highest accuracy, underscoring the benefits of equal, collaborative interaction among agents. Our dialogue act analysis revealed that specific communicative behaviors, such as statements (S), acknowledge (ACK), and hints (H), play pivotal roles in the agents' collaborative reasoning process. These findings offer valuable insights into how different interaction patterns contribute to effective problem

solving and provide a foundation for designing more adaptive AI tutoring systems. This research is ongoing. Future efforts will concentrate on (i) strengthening agent-to-agent communication protocols, (ii) applying advanced prompt-optimization techniques, and (iii) experimenting with heterogeneous agent configurations to further improve performance, and (iv) integrating additional LLMs (e.g., DeepSeek and Llama series). We also plan a deeper, systematic analysis of dialogue acts and their specific roles in collaborative problem solving. Overall, our work demonstrates that structured, multi-agent communication not only boosts accuracy in mathematical problem-solving but also enriches the underlying dialogue dynamics, paving the way for more effective AI educational platforms.

Acknowledgments. The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C240010 to University of Georgia Research Foundation. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

1. Justin, C.-Y.C., Saha, S., Bansal, M.: Reconcile: round-table conference improves reasoning via consensus among diverse llms. arXiv preprint arXiv:2309.13007 (2023)
2. Hagendorff, T.: Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988 (2023)
3. Hendrycks, D., et al.: Measuring mathematical problem solving with the MATH dataset. In: NeurIPS (2021)
4. Hurst, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
5. Keating, V.: Zero-shot mathematical problem solving with large language models via multi-agent conversation programming. In: AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI (2024)
6. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1-2), 81–93 (1938)
7. Latif, E., et al.: A systematic assessment of openai o1-preview for higher order thinking in education. arXiv preprint arXiv:2410.21287 (2024)
8. Li, H., et al.: Theory of mind for multi-agent collaboration via large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 180–192. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.13. https://aclanthology.org/2023.emnlp-main.13/
9. Liang, T., et al.: Encouraging divergent thinking in large language models through multi-agent debate. arXiv: 2305.19118 (2024)
10. Liang, Z., et al.: MathChat: benchmarking mathematical reasoning and instruction following in multi-turn interactions. arXiv preprint arXiv:2405.19444 (2024)
11. Lin, J., et al.: Is it a good move? Mining effective tutoring strategies from human-human tutorial dialogues. Futur. Gener. Comput. Syst. **127**, 194–207 (2022)

12. Lin, J., et al.: Robust educational dialogue act classifiers with low resource and imbalanced datasets. In: International Conference on Artificial Intelligence in Education, pp. 114–125. Springer, Heidelberg (2023)
13. Tan, W., et al.: Does informativeness matter? Active learning for educational dialogue act classification. In: International Conference on Artificial Intelligence in Education, pp. 176–188. Springer, Heidelberg (2023)
14. Vail, A.K., Boyer, K.E.: Identifying effective moves in tutoring: on the refinement of dialogue act annotation schemes. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 199–209. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07221-0_24
15. Van Lehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Educ. **16**(3), 227–265 (2006)
16. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes, vol. 86. Harvard university press (1978)
17. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
18. Yu, J., et al.: From mooc to maic: reshaping online teaching and learning through llm-driven agents. arXiv preprint arXiv:2409.03512 (2024)
19. Yue, M., et al.: Mathvc: an llm-simulated multi-character virtual classroom for mathematics education. arXiv preprint arXiv:2404.06711 (2024)
20. Zhang, L., et al.: Data augmentation for sparse multidimensional learning performance data using generative AI. IEEE Trans. Learn. Technol. (2025)
21. Zhang, L., et al.: Exploring communicative strategies for dual LLM agents in mathematical problem solving (2022)
22. Zhang, L., et al.: Predicting learning performance with large language models: a study in adult literacy. In: International Conference on Human-Computer Interaction, pp. 333–353. Springer, Heidelberg (2024). https://doi.org/10.1007/978-3-031-60609-0_24
23. Zhang, L., et al.: SPL: a socratic playground for learning powered by large language mode. arXiv preprint arXiv:2406.13919 (2024)



VETTING AI for Deeper Learning: Constraining LLMs to Encourage Student Inquiry

Shan Zhang(✉), Hongming Li, Seiyon M. Lee, Noah L. Schroeder, and Anthony F. Botelho

University of Florida, Gainesville, FL 32601, USA
{zhangshan,hli3,leeseiyon,schroedern,a.botelho}@ufl.edu

Abstract. As generative AI tools like ChatGPT are integrated into education, concerns persist about over-reliance, where students seek direct answers without engaging in critical thinking. To address this, we propose VETTING, a framework for designing pedagogically-aligned applications of large language models (LLMs) in education. To explore the practical implementation of this framework, we present a case study examining the impact of a VETTING-informed chatbot that includes a verification layer designed to restrict the provision of direct problem solutions to students. In a randomized trial with 41 undergraduate students, participants interacted with either a general GPT-4o model or a VETTING-informed version designed to promote problem-solving. Findings indicate that students engaged with AI in unexpected ways, rarely asking direct questions. Instead, they sought clarification or broke problems into smaller parts. They treated the chatbot as an extension of their learning materials and rarely followed up when direct answers were unavailable. While students in the VETTING group interacted more frequently, their engagement lacked sustained inquiry. These findings highlight the complexities of designing AI tools that foster deeper learning, but offer insights into how frameworks like VETTING can be implemented to promote targeted pedagogical practices.

Keywords: Large Language Models (LLMs) · AI Literacy · AI Framework · AI Chatbot

1 Introduction

The integration of large language models (LLMs) like GPT into educational contexts presents significant opportunities to enhance student learning and engagement [18]. Educators and researchers have explored various strategies to leverage LLMs' potential while ensuring alignment with sound pedagogical principles and responsible AI practices [5]. Despite the many potential advantages LLMs can offer [1,3,5,13,15], concerns about student over-reliance on LLMs have

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 266–273, 2025.
https://doi.org/10.1007/978-3-031-99264-3_33

gained increasing attention as these models become more common in educational settings.

Over-reliance occurs when students depend excessively on LLM-generated responses without questioning or critically evaluating their accuracy [10]. This can lead to increased errors and a decline in independent problem-solving abilities [2]. Research has shown that students' trust in AI responses and their ability to assess the reliability of AI-generated content are key factors influencing over-reliance [8]. These are key aspects of AI literacy [11], which is a construct of growing importance as it can influence how people use and learn with AI.

Given these concerns, it is critical to help learners develop a clearer understanding of how to effectively use LLMs in learning settings. One way we can do this is by enhancing their AI literacy to foster meaningful engagement with AI tools. To mitigate over-reliance specifically, we propose the Verification and Evaluation Tool for Targeting Invalid Narrative Generation (VETTING) framework, which is designed to reshape how students interact with AI in learning environments. By implementing VETTING in a classroom setting through a case study, we specifically investigate:

RQ1. How do students using the VETTING-informed GPT model differ from those using the general GPT model in terms of academic performance, AI literacy, and engagement behaviors?

RQ2. What insights can be gained from the design and implementation of a chatbot with a verification layer to support student learning?

2 VETTING Framework

VETTING incorporates a dual-stage verification process, consisting of input pre-processing and output post-validation, to provide an iterative feedback loop rather than a binary accept or reject mechanism as is used in classification-based approaches such as Llama Guard [9]. This process is embedded within a three-layer architecture comprising the User Interaction Layer, Verification Middleware, and Language Model Integration Layer. Further details on each component are provided below.

Three-Layer Architecture. The system is structured into three layers: (A) User Interaction (UI), (B) Verification Middleware (VM), and (C) Language Model Integration (LMI). The UI layer serves as the student-facing interface, capturing queries, presenting filtered responses, maintaining session context, and applying educational scaffolding. This layer can function as any LLM-based pedagogical agent designed as a chatbot. The VM layer is the core of the framework, where input pre-processing, output validation, and post-processing occur. It detects and redirects ineffective learning strategies, such as immediate answer-seeking, while promoting productive help-seeking behaviors and scaffolded learning. The LMI layer manages prompt engineering and response generation, ensuring that outputs align with verification feedback while maintaining conversational coherence and pedagogical relevance.

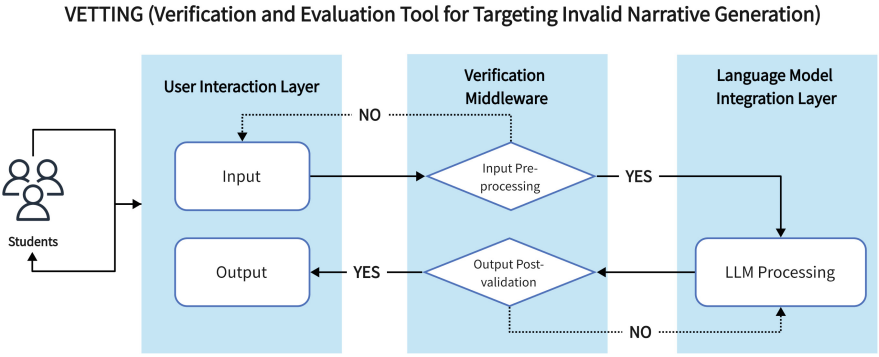


Fig. 1. The VETTING framework's three-layer architecture.

Stage 1. Input Pre-processing. This stage operates within the VM, analyzing student queries through semantic pattern recognition and contextual analysis. The filtering mechanism identifies direct answer-seeking behaviors while preserving meaningful inquiry patterns, balancing student intent with educational objectives. As illustrated in Fig. 1, VETTING processes student input to detect potentially undesirable requests. Examples include attempts to manipulate the LLM's behavior (e.g., re-prompting for specific responses) or unproductive strategies (e.g., directly requesting answers).

Stage 2. Output Post-validation Stage. This stage implements a recursive verification mechanism, where each AI-generated response undergoes multi-criteria evaluation before being delivered to the student. As shown in Fig. 1, VETTING captures the LLM-generated response before displaying it to the student, verifying its alignment with content guidelines. If the validation process detects restricted content, such as direct answers to assessment questions, it initiates an iterative refinement process, modifying the response or re-prompting the LLM until it meets content guidelines. If the response fails to meet the criteria within a set number of attempts, the system ends the refinement and informs the student that a direct answer cannot be provided and offers targeted scaffolding, such as rephrasing suggestions, problem decomposition, or related concept exploration.

3 Case Study

To examine the effect of VETTING on how students' interaction with LLM and learning, we conducted a randomized controlled study through VETTING Chat, a web-based AI chatbot building on the VETTING framework. While VETTING supports both input pre-processing and output post-validation, this paper focuses on output verification which includes Verification LLM for filtering

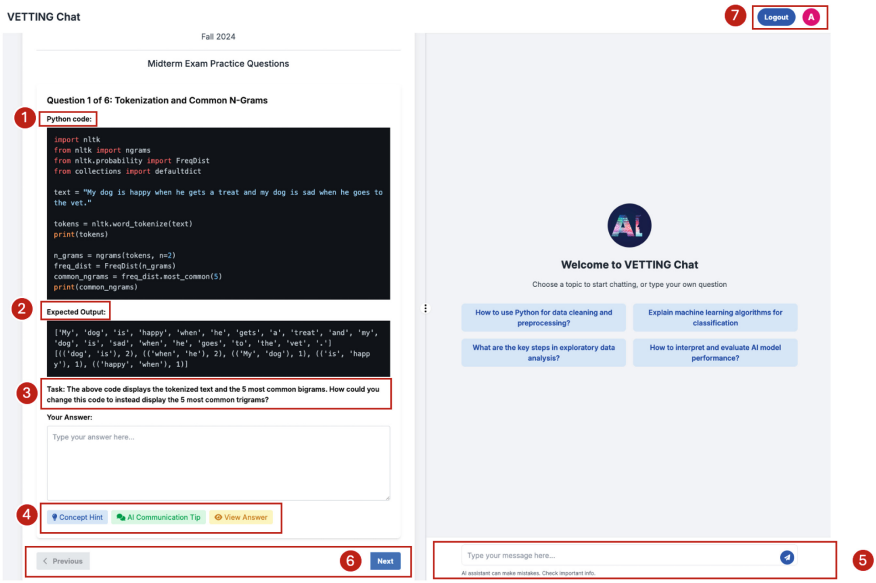


Fig. 2. A VETTING-informed Chatbot's dual-panel design with learning materials (left) and AI chat interface (right).

Chat LLM's response to students and Chat LLM for student response as part of our initial deployment stage[1].

Participants were undergraduate students enrolled in an introductory machine learning (ML) and AI course at a large southeastern U.S. university in Fall 2024. As an open course to students of any major, 70 students from 10 colleges were enrolled. Students were invited to participate in a voluntary midterm practice exam. The practice exam was administered through a web interface (Fig. 2). Forty one students participated in the study at a location of their choosing (e.g., home, school, etc.) and were randomly assigned to one of two conditions: (1) the VETTING-informed GPT-4o condition, where the model provided structured guidance without direct answers, or (2) the general GPT-4o condition[2].

The study was divided into three phases. In the first phase, participants completed six midterm practice questions that covered basic ML concepts while interacting with their assigned GPT condition. Following this, they proceeded to a post-test phase, where they answered three post-test questions without access

[1] Both models in this study used GPT-4o; detailed prompt design for both models are available on OSF: https://osf.io/fyz3e/?view_only=5cbf1b35f29e4e6cac92e503b5dad052.

[2] This study was conducted under an IRB-approved protocol and was pre-registered on OSF, which has been blinded for review.

to GPT. The study concluded with a 30-minute AI literacy test, which included 30 multiple-choice questions and one sorting question

Four types of data were collected: (1) midterm practice exam responses, (2) post-test answers, (3) an AI literacy test, and (4) backend interaction logs, including GPT conversations, query logs, response patterns, and click-stream data. The AI literacy test is adapted from existing literature [7] that researchers developed and validated items with undergraduate students similar to our context and mapped each item with an AI literacy theoretical framework [11].

A mixed-methods approach was used to analyze the data. First, non-parametric MannâĂŞWhitney U tests were conducted to examine whether there were significant differences between the VETTING and general GPT groups in post-test performance, AI literacy scores, and engagement metrics. Engagement metrics included total problem visits, total AI communication tip clicks, total concept hint clicks, and total view answer clicks. These comparisons were conducted to assess group-level differences without assuming normality, given the small sample size and data distribution characteristics. Second, two authors used an inductive thematic analysis [4] to develop a coding scheme and coded interactions between students and the chatbot for two groups.

4 Results

Not all of the participants in the study completed all of the tasks. In the VETTING group, 14 completed all tasks, with 12 finishing the AI literacy test. In the general group, 8 completed all tasks, and 6 finished the AI literacy test. High attrition was likely due to the study's optional nature and remote format.

To address RQ1, we examined whether students in the VETTING and general GPT groups differed in post-test performance, AI literacy, or engagement metrics. Mann-Whitney U tests revealed no significant differences in performance ($\text{Median}_{\text{VETTING}} = 4.5$, $\text{Median}_{\text{General}} = 4$, $U = 72$, $p = 0.278$), AI literacy Scores ($\text{Median}_{\text{VETTING}} = 20.5$, $\text{Median}_{\text{General}} = 19.5$, $U = 42.5$, $p = 0.573$), or engagement metrics: Total Concept Hint Clicks ($\text{Median}_{\text{VETTING}} = 1.5$, $\text{Median}_{\text{General}} = 1.5$, $U = 52.5$, $p = 0.833$), AI Tips ($\text{Median}_{\text{VETTING}} = 0.5$, $\text{Median}_{\text{General}} = 0.5$, $U = 56.5$, $p = 1$), Total View Answer Clicks ($\text{Median}_{\text{VETTING}} = 4.5$, $\text{Median}_{\text{General}} = 4$, $U = 57$, $p = 0.972$), and Total Problem Visits ($\text{Median}_{\text{VETTING}} = 9$, $\text{Median}_{\text{General}} = 7.5$, $U = 62.5$, $p = 0.661$).

In terms of student-LLM interaction patterns (RQ2), a total of 218 chat messages were exchanged between students. Surprisingly, only three students overall triggered an LLM response that would have contained a direct answer, requiring revision through VETTING's verification process. Upon further analysis, in the VETTING group, 12 students interacted with the chatbot, generating 42 queries, while in the general group, 8 students also asked a total of 42 questions. The full result table is available on OSF[3].

[3] Qualitative labeling is available on OSF: https://osf.io/fyz3e/?view_only=5cbf1b35f29e4e6cac92e503b5dad052.

We found that students interacted with the AI in ways that diverged from our initial expectations. They rarely used direct queries such as "Can you show me/tell me/send me?" This challenged our assumptions about the types of questions students would ask, leading to cases where the verification layer failed to detect Chat LLM responses that included direct answers. In addition, students often treated the chatbot as an extension of their learning materials, frequently copying and pasting task questions or breaking them into smaller components. When answering questions, they didn't just copy and paste what they got from the model, instead, they tended to paraphrase AI responses in their own words, indicating some critical engagement. While students in the VETTING group engaged in more, they showed limited effort in seeking alternative explanations when direct answers were unavailable.

5 Discussion

To better align the application of LLMs with pedagogical approaches, several companies have recently developed systems specifically designed to address pedagogical gaps and better support student learning. For example, recently introduced systems such as LearnLM and Claude for Education aim to promote student problem-solving ability and critical thinking, rather than relying on simple answer-giving. LearnLM incorporates pedagogical objectives into its training pipeline, combining supervised fine-tuning and reinforcement learning from human feedback to encourage models to follow instructions like "don't give away the answer" during educational dialogues [16]. Similarly, Claude's new Learning Mode applies Socratic prompting techniques and conceptual guidance to scaffold reasoning and critical thinking. However, both are early-stage efforts with limited evidence of their ability to consistently prevent direct answer-giving. While both systems incorporate instruction-tuned prompting, Socratic questioning, VETTING adds a unique post-response verification layer that explicitly blocks direct answers, providing a more robust approach to enforcing inquiry-driven interactions. In fact, this need for post-response control aligns with recommendations from Nye et al. [14] who emphasize the importance of the external validation of LLMs to ensure pedagogical alignment in dialog-based tutoring systems.

Our research offers preliminary insights into student-AI interaction patterns. It also highlighted occasional inconsistencies in LLM behavior, where both Verification LLM and Chat LLM failed to follow instructions and provided direct answers despite explicit prompts not to. This failure of LLMs to follow instructions has been documented elsewhere as well [12,19]. While we understand that instruction-following alone may not be reliable and thus requires post-response validation, the limited effectiveness of the current verification layer in handling diverse query patterns that challenge our assumptions suggests the need for more sophisticated approaches to model alignment and output control such as few-shot learning [6], multi-turn dialogue management [17] and adding input pre-processing through ML-LLM from VETTING. These findings highlight the complexities of integrating pedagogically aligned AI into learning environments

and emphasize the importance of designing tools that consider varying degrees of students' AI literacy and actively support meaningful student engagement rather than passive reliance.

Acknowledgments. We thank the National Science Foundation (#2331379), Bill and Melinda Gates Foundation, Schmidt Futures, and OpenAI for supporting this work.

References

1. Amira Roumaissa Boudjedra, H.E.K.: Promoting EFL learner's self-regulated learning through the use of artificial intelligence applications (2024)
2. Bo, J.Y., Wan, S., Anderson, A.: To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. arXiv preprint arXiv:2412.15584 (2024)
3. Chang, D.H., Lin, M., Hajian, S., Wang, Q.Q.: Educational design principles of using AI chatbot that supports self-regulated learning in education: goal setting, feedback, and personalization. Sustainability **15**(17), 12921 (2023)
4. Clarke, V., Braun, V.: Thematic analysis. J. Posit. Psychol. **12**(3), 297–298 (2017)
5. Fu, Y., Weng, Z.: Navigating the ethical terrain of AI in education: a systematic review on framing responsible human-centered AI practices. Comput. Educ. Artif. Intell. 100306 (2024)
6. Gharoun, H., Momenifar, F., Chen, F., Gandomi, A.H.: Meta-learning approaches for few-shot learning: a survey of recent advances. ACM Comput. Surv. **56**(12), 1–41 (2024)
7. Hornberger, M., Bewersdorff, A., Nerdel, C.: What do university students know about artificial intelligence? Development and validation of an AI literacy test. Comput. Educ. Artif. Intell. **5**, 100165 (2023)
8. Hunter, R., Moulange, R., Bernardi, J., Stein, M.: Monitoring human dependence on AI systems with reliance drills. arXiv preprint arXiv:2409.14055 (2024)
9. Inan, H., et al.: Llama guard: LLM-based input-output safeguard for human-AI conversations. arXiv preprint arXiv:2312.06674 (2023)
10. Korpimies, K., Laaksonen, A., Luukkainen, M.: Unrestricted use of LLMs in a software project course: student perceptions on learning and impact on course performance. In: Proceedings of the 24th Koli Calling International Conference on Computing Education Research, pp. 1–7 (2024)
11. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2020)
12. Lou, R., Zhang, K., Yin, W.: A comprehensive survey on instruction following. arXiv preprint arXiv:2303.10475 (2023)
13. Min, T., Lee, B., Jho, H.: Integrating generative artificial intelligence in the design of scientific inquiry for middle school students. Educ. Inf. Technol. 1–32 (2025)
14. Nye, B.D., Mee, D., Core, M.G.: Generative large language models for dialog-based tutoring: an early consideration of opportunities and concerns. In: LLM@ AIED, pp. 78–88 (2023)
15. Sahab, S., Haqbeen, J., Ito, T.: Conversational AI as a facilitator improves participant engagement and problem-solving in online discussion: sharing evidence from five cities in Afghanistan. IEICE Trans. Inf. Syst. **107**(4), 434–442 (2024)

16. Team, L., et al.: Learnlm: improving gemini for learning. arXiv preprint arXiv:2412.16429 (2024)
17. Yi, Z., Ouyang, J., Liu, Y., Liao, T., Xu, Z., Shen, Y.: A survey on recent advances in LLM-based multi-turn dialogue systems. arXiv preprint arXiv:2402.18013 (2024)
18. Zhao, R., Yunus, M.M., Rafiq, K.R.M., et al.: The impact of the use of chatgpt in enhancing student's engagement and learning outcomes in higher education: a review. Int. J. Acad. Res. Bus. Soc. Sci. **13**(12) (2023)
19. Zhou, J., et al.: Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911 (2023)

Exploring the Feasibility of Personalized AI Feedback to Improve Children's Planning Skills

Somphop Sukjaitham[1,3(✉)], Garvin Brod[1,2,3], and Jasmin Breitwieser[1,3]

[1] Education and Human Development, DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany
{s.sukjaitham,g.brod,j.breitwieser}@dipf.de
[2] Department of Psychology, Goethe University, Frankfurt am Main, Germany
[3] IDeA-Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany

Abstract. Making plans that specify when and where to perform goal-directed actions—known as implementation intentions—is a powerful self-regulation strategy that promotes goal achievement, particularly when the plans are well-formulated. However, children in particular struggle to formulate plans that meet the criteria to be considered effective. Therefore, the present study investigates whether supportive feedback can help improve the quality of their self-created plans. Specifically, we test the feasibility of employing generative AI to deliver personalized feedback tailored to each child's input and compare its effectiveness to that of generic, non-personalized feedback created by experts. We plan to recruit 80 children aged 10 to 13 who will participate in a within-subject experimental study. Each child will complete six planning trials in which they first generate a plan, receive either personalized AI feedback or generic feedback, and then revise their plan. We will test whether plan quality improves after AI feedback, and whether AI feedback leads to greater plan quality improvement than high-quality generic feedback. In addition, we examine whether children perceive AI feedback as more helpful and motivating compared to generic feedback. The research findings are expected to (1) advance our understanding of how AI-generated feedback influences children's planning skills, (2) provide deeper understandings of children's perceptions on AI-generated output, and (3) inform the design of adaptive feedback systems in education.

Keywords: AI · Feedback · Self-Regulated Learning · Implementation Intentions

1 Introduction

Children often encounter a variety of learning challenges, such as difficulty initiating tasks due to a lack of motivation [1], maintaining consistent study habits [2], or overcoming procrastination [3]. To address these challenges, research on self-regulated learning (SRL) highlights the importance of planning in supporting students' goal setting, action organizing and progress monitoring skills [4, 5]. However, children often struggle with

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 274–281, 2025.
https://doi.org/10.1007/978-3-031-99264-3_34

planning, which limits their ability to regulate their learning effectively. One efficient strategy to support planning is the use of implementation intentions, or if-then plans, which link specific situations to goal-directed actions and help automate responses to cues [6]. For instance, "If I get back home from school, then I will study vocabulary" connects the situational cue (getting back home from school) to the intended action (studying vocabulary). Research has shown that the effectiveness of plans depends on their quality, especially the situational cue [6, 7]. Plans with cues that are specific, externally anchored, salient, appropriate, and temporally precise are more likely to lead to action, however children often formulate substandard plans [8, 9]. To take up this issue, the present study builds on our research-based mobile app that supports children's planning skills via if-then planning technique. Yet this context raises the vital question of how children can assess and improve their plans. They need timely, individualized feedback to achieve this, which traditional methods often lack due to delays, generalization, or classroom constraints [9–11]. With the advancement of technology, personalized AI feedback offers a promising alternative by providing instant, tailored guidance that can enhance the quality of children's self-formulated plans.

Recognizing this potential, the advent of artificial intelligence (AI) applications in education has introduced new opportunities for developing personalized feedback systems. Although AI-generated feedback has demonstrated effectiveness in enhancing learning and skills development (e.g., [12–16]), research on its impact in supporting children's planning skills is still lacking. This raises an essential question of whether AI-generated feedback offers advantages beyond conventional feedback systems or instructional scaffolding. This study aims to make three key contributions to research in AI in Education. First, it provides evidence-based insights on the effectiveness of AI-generated feedback in improving planning skills, emphasizing how children enhance their self-created if-then plans by comparing personalized AI feedback with a strong control condition—high quality generic feedback. While both types of feedback include explanation based on characteristics of good if-then plans, personalized AI feedback offers individualized components tailored to children's inputs. Second, it contributes to a deeper understanding of children's perceptions of AI-generated output, particularly in relation to technology acceptance. It explores how children perceive personalized AI feedback differently from expert-generated generic feedback in terms of helpfulness and motivational impact. Lastly, this research provides practical implications for the design of adaptive AI-powered educational tools which can be adapted to other types of educational technology.

To accomplish these contributions, this study is anchored on three primary research questions. The first question is whether personalized AI feedback helps children to improve the quality of their plans. The second question is whether personalized AI feedback is more effective for improving plan quality than non-personalized expert feedback. Another direction of research asks the question whether children perceive personalized AI feedback as more useful than generic feedback. To explore this more precisely, the study examines two crucial aspects: (1) perceived helpfulness and (2) perceived motivational impact. By shedding light on these questions, this study will provide insights into the optimal design of personalized AI feedback to ensure that it effectively supports learning and engagement in educational contexts. The present study

builds on pilot data and represents a pre-registered experiment that will be conducted before the conference. The pilot phase served to refine feedback prompts, study materials, and procedures, ensuring methodological robustness for the main study.

2 Methodology

2.1 Participants

The experiment will be conducted with children aged 10 to 13 years, recruited through collaboration with a local school. The participants include children of mixed genders who are fluent in German, which is the language used in the study. The target sample size of 80 participants was determined through a simulation-based power analysis using Monte Carlo simulations for ordinal regression models (see [17]).

2.2 Experimental Design

In this study, individuals will engage in a planning task where they formulate if-then plans, receive feedback, and revise their plans accordingly. After completing each trial, they will also rate their perception of the feedback, assessing its helpfulness and motivational impact. They will construct only the if-part of an if-then plans, specifying a situational cue that prompts goal-directed behavior. This study adopts a within-subject experimental design with two independent variables—feedback type and time. The feedback type consists of two levels, personalized AI and generic feedback, while time consists of pre-feedback and post-feedback conditions. Each participant will encounter both types of feedback in a randomized sequence across six trials. Each participant will receive three trials assigned to generic feedback and three to personalized AI feedback, ensuring equal exposure. To minimize expectancy effects, participants will be blinded to the differences between feedback conditions across trials.

Personalized AI feedback is generated by ChatGPT-4 by systematic prompts and provides individualized feedback tailored to each child's if-then plans based on predefined quality criteria. This type of feedback aligns with the child's input, ensuring alignment with their specific needs. In contrast, generic feedback consists of pre-written responses. However, we tried to make this control condition as strong as possible by a) having experts on implementation intentions create the responses based on the same quality criteria as given to ChatGPT and including a motivational statement, and b) having five different versions of generic feedback so that children do not receive the exact same feedback message twice, to ensure that generic feedback is not perceived as boring. For each trial, children will initially formulate a plan independently, serving as the pre-feedback measure. After receiving either personalized or generic feedback, they will have the opportunity to revise their plan, representing the post-feedback measure.

To reduce order effects, randomization occurs at the trial level, where each participant experiences both feedback conditions in a random sequence. Additionally, six different scenarios are randomly assigned across trials to introduce contextual variation. These scenarios are designed to reflect common tasks relevant to children, encompassing school-related activities, household responsibilities, and hobbies. Specifically, they include (1) learning to program, (2) writing stories, (3) drawing pictures, (4) walking a dog, (5) shooting videos, and (6) watering houseplants.

2.3 Measured Variables

The primary dependent variable is planning quality, evaluated using a rubric-based scoring system adapted from Schaaf et al. [9]. Plans will be assessed on a four-point scale based on five key dimensions. A score of one indicates a high-quality cue that is specific, external, salient, appropriate, and temporally precise. A score of two represents a partially effective cue that lacks one or more of these characteristics. A score of three represents a weak cue that either lacks a situational cue entirely or relies on internal cues. A score of four reflects an invalid response that fails to provide a meaningful plan. Improvement in plan quality will be measured as the difference between pre-feedback and post-feedback scores, capturing the main effect receiving feedback. To ensure reliability, two independent researchers will evaluate the plans and report on inter-rater reliability. Feedback conditions will be blinded during rating to curtail potential bias.

The second dependent variable is perceived usefulness of feedback, assessed after each trial using a four-point Likert scale. Two questions will measure children's evaluations of the feedback. Perceived helpfulness will be assessed by asking, "How helpful did you find feedback in improving your plan?" with response options ranging from very helpful to not at all. The perceived motivation will be measured by asking, "How much did the feedback motivate you to revise your plan?" with response options ranging from very motivating to not at all.

2.4 Data Collection and Planned Analysis

In the experiment, children will interact with a web-based feedback system to create if-then plans and receive feedback according to the assigned condition in each trial. The study follows a structured procedure designed to guide children through the task. At the beginning of the session, the children are informed about the purpose of learning and the importance of planning. They then complete a short pre-task survey and are given a brief introduction to if-then planning. Once children understand the concept, they are introduced to the study format through a brief explanation of the number of trials and the estimated time required per trial and a short video walkthrough of the web-based feedback system, explaining how it works and how feedback is delivered. Before beginning the main trials, children complete a practice trial to familiarize themselves with the task. In the main trials, children complete six planning tasks, each introduced through a brief scenario description that includes a predefined goal to provide context for the planning task.

For full analysis, an ordinal multilevel logistic regression model will be employed to investigate whether children show greater improvement in plan quality after receiving personalized AI feedback compared to generic feedback. The model will include feedback type (personalized AI feedback vs. generic feedback), time (pre-feedback vs. post-feedback), and their interaction term as predictors. A random intercept for participant ID (a unique identifier assigned to each participant) will be included to account for within-subject dependencies. In addition, a paired-samples t-test is expected to be conducted on the perceived helpfulness and perceived motivation scores to compare perceived usefulness of feedback between conditions.

3 Preliminary Insights

As part of the study development, an initial data collection phase was conducted to examine how AI-generated feedback influences children's planning skills and their perceptions of feedback. Note that this phase served as a pilot study; data collection for the full study with 80 participants will be completed before the conference. Insights from this pilot phase provide preliminary evidence on the effectiveness of personalized AI feedback while also revealing individual differences in how children respond to feedback interventions. The preliminary analysis was instrumental in refining feedback prompts, study materials, and procedures, ensuring that participants clearly understood the task and that the selected methods effectively captured the intended measurements.

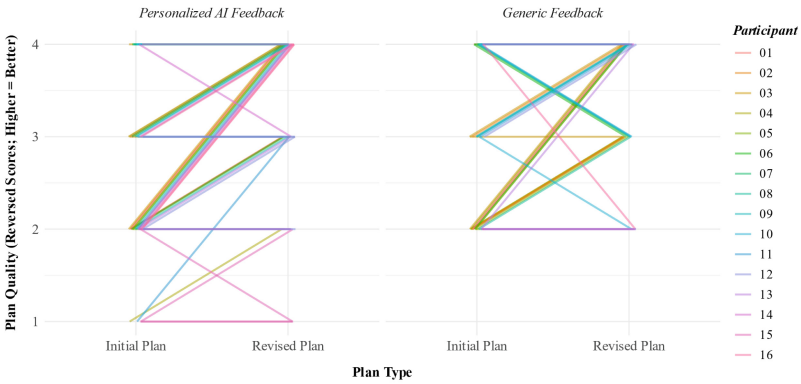


Fig. 1. Individual changes in plan quality before and after feedback across feedback conditions

A preliminary analysis of children's planning performance suggests that personalized AI feedback may contribute to improvements in plan quality. Initial data, collected from $N = 16$ children who completed 96 plans, indicates that personalized AI feedback led to greater improvement in plan quality compared to high-quality generic feedback.

As shown in Fig. 1, after receiving personalized AI feedback, average plan quality scores increased from $M = 2.62$ $(SD = 0.97)$ before feedback to $M = 3.43$ $(SD = 0.92)$. After receiving high-quality generic feedback, plan quality scores increased from $M = 2.77$ $(SD = 0.87)$ to $M = 3.23$ $(SD = 0.81)$. Although these preliminary results indicate that feedback may contribute to plan quality improvement, individual differences in responsiveness to feedback were observed. Some plans after feedback demonstrated substantial improvements, while others exhibited minimal change, suggesting variability in how children process and apply feedback.

From another angle, children's perceptions of personalized AI feedback were generally positive, with a majority rating the feedback as helpful or very helpful. On average, children rated personalized AI feedback as more useful $(M = 2.17, SD = 0.85)$ and motivating $(M = 2.40, SD = 0.77)$ compared to generic feedback, which received lower ratings for usefulness $(M = 1.58, SD = 1.05)$ and motivation $(M = 1.74, SD = 0.93)$

even though it was generated by experts. These findings suggest that children may find personalized AI feedback more useful in their learning process.

4 Challenges and Future Works

This study will provide preliminary insights into the effectiveness of personalized AI feedback in enhancing children's planning skills. The sample size limitation remains a key challenge at this stage, emphasizing the need to recruit the targeted number of participants to ensure sufficient statistical power. The pre-registered data collection will start soon and is expected to be completed by the time of the conference.

Based on the results of this proof-of-concept study, our future research will explore the long-term effects of personalized AI feedback by conducting longitudinal studies that examine how repeated exposure to AI-generated feedback influences children's planning skills over time. Further research could also extend beyond controlled experimental settings by conducting real-world field studies, testing personalized AI feedback in authentic learning environments. Additionally, replicating this study across different target skills, content domains, or educational levels would help establish the broader applicability of AI-generated feedback in diverse learning contexts.

5 Conclusion

This study explores the feasibility of AI-generated personalized feedback in improving children's planning skills. Preliminary findings suggest that children may demonstrate greater improvement in plan quality when receiving personalized AI feedback compared to generic feedback. In addition, children appear to perceive personalized AI feedback as more useful, both in terms of helpfulness and motivation. These findings contribute to ongoing discussions on the role of artificial intelligence in fostering personalized learning, supporting the development of more responsive and adaptive educational environments. By tailoring feedback to individual learners, AI-powered systems can provide more meaningful guidance, promoting autonomy and self-regulated learning skills. These insights also inform the design and development of AI-driven educational tools, ensuring they align with learners' needs and enhance engagement. Moreover, understanding children's perceptions of AI in relation to technology acceptance is crucial for fostering trust in AI-assisted learning environments. This study lays the groundwork for further investigations into the long-term effects of personalized AI feedback through longitudinal studies, real-world field studies, and replications across different educational contexts.

Acknowledgments. This study was supported by a Leibniz Transfer grant from the Leibniz Association and by the CERES network (Connecting the EdTech Research Ecosystem). The authors thank the participating schools, students, and teachers for their support. We thank Anawat Tongta for his invaluable guidance in developing the study materials and Mirijam Schaaf for productive discussions on the plan quality rubric.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Wolters, C.A.: Regulation of motivation: evaluating an underemphasized aspect of self-regulated learning. Educ. Psychol. **38**(4), 189–205 (2003). https://doi.org/10.1207/S15326985EP3804_1
2. Winne, P.H., Hadwin, A.F.: The weave of motivation and self-regulated learning. In: Schunk, D.H., Zimmerman, B.J. (eds.) Motivation and Self-Regulated Learning: Theory, Research, and Applications, pp. 297–314. Lawrence Erlbaum Associates Publishers, Mahwah (2008)
3. Steel, P.: The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. Psychol. Bull. **133**(1), 65–94 (2007). https://doi.org/10.1037/0033-2909.133.1.65
4. Zimmerman, B.J.: Becoming a self-regulated learner: an overview. Theory Pract. **41**(2), 64–70 (2002). https://doi.org/10.1207/s15430421tip4102_2
5. Zimmerman, B.J., Moylan, A.R.: Self-regulation: where metacognition and motivation intersect. In: Hacker, D.J., Dunlosky, J., Graesser, A.C. (eds.) Handbook of Metacognition in Education, pp. 299–315. Routledge/Taylor & Francis Group, New York (2009)
6. Gollwitzer, P.M.: Implementation intentions: strong effects of simple plans. Am. Psychol. **54**(7), 493–503 (1999). https://doi.org/10.1037/0003-066X.54.7.493
7. Sheeran, P., Listrom, O., Gollwitzer, P.M.: The when and how of planning: meta-analysis of the scope and components of implementation intentions in 642 tests. Eur. Rev. Soc. Psychol. **36**, 1–33 (2024). https://doi.org/10.1080/10463283.2024.2334563
8. Greene, J.A., Hutchison, L.A., Costa, L.-J., Crompton, H.: Investigating how college students' task definitions and plans relate to self-regulated learning processing and understanding of a complex science topic. Contemp. Educ. Psychol. **37**(4), 307–320 (2012). https://doi.org/10.1016/j.cedpsych.2012.02.002
9. Schaaf, M., Brod, G., Breitwieser, J.: Planning promotes studying – the higher the plan quality the better: a micro-randomized trial (2024). https://doi.org/10.31234/osf.io/jsmq4
10. Nicol, D.J., Macfarlane-Dick, D.: Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Stud. High. Educ. **31**, 199–218 (2006). https://doi.org/10.1080/03075070600572090
11. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007). https://doi.org/10.3102/003465430298487
12. Olivera-Aguilar, M., Lee, H.-S., Pallant, A., Belur, V., Mulholland, M., Liu, O.L.: Comparing the effect of contextualized versus generic automated feedback on students' scientific argumentation. ETS Res. Rep. Ser. **2022**, 1–14 (2022). https://doi.org/10.1002/ets2.12344
13. Wynn, A.T., Wang, J., Umezawa, K., Cristea, A.I.: An AI-based feedback visualisation system for speech training. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium, pp. 510–514. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11647-6_104
14. Dai, W., et al.: Can large language models provide feedback to students? A case study on ChatGPT. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 323–325 (2023). https://doi.org/10.1109/ICALT58122.2023.00100
15. Jacobsen, L.J., Weber, K.E.: The Promises and Pitfalls of LLMs as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback (2023). https://doi.org/10.31219/osf.io/cr257

16. Gombert, S., et al.: From the automated assessment of student essay content to highly informative feedback: a case study. Int. J. Artif. Intell. Educ. **34**, 1378–1416 (2024). https://doi.org/10.1007/s40593-023-00387-6
17. Gambarota, F., Altoè, G.: Ordinal regression models made easy: a tutorial on parameter interpretation, data simulation and power analysis. Int. J. Psychol. **59**, 1263–1292 (2024). https://doi.org/10.1002/ijop.13243



Exploring AI Chatbots for Learning Mathematics: Students' Perspectives on Accuracy and Educational Value

Jacek Stańdo[1], Żywilla Fechner[1(✉)], Anita Dąbrowicz-Tlałka[2], Katarzyna Kujawska[2], and Magdalena M. Musielak[2]

[1] Lodz University of Technology, Lodz, Poland
{jacek.stando,zywilla.fechner}@p.lodz.pl
[2] Gdansk University of Technology, Gdansk, Poland
{anita,kkujawsk,magmusie}@pg.edu.pl

Abstract. This study explores the use of artificial intelligence (AI) chatbots in mathematics education. With the increasing integration of AI tools such as ChatGPT in higher education, particularly in engineering courses, this research investigates whether students can identify errors in AI-generated solutions. Data was collected from over 200 s-semester engineering students at Gdańsk University of Technology, who evaluated mathematical problems solved by three different AI chatbots, assessing the correctness, difficulty, and usefulness of the AI solutions for self-study. The results revealed that students consistently rated AI-generated solutions higher than those provided by human teachers, with significant differences in evaluations. This suggests a potential challenge in students' ability to critically evaluate AI responses, possibly due to overreliance on the technology. Additionally, a high school-level mathematics task from the Polish Matura exam was analyzed as part of the study, providing further insight into AI's performance across different educational levels. The findings highlight the importance of fostering critical thinking skills in students, ensuring they can effectively integrate AI into their learning while maintaining the ability to question and verify AI-generated results. As AI continues to evolve, the study underscores the need for educational strategies that balance technology integration with the development of human mathematical and analytical skills.

Keywords: AI in maths education · AI-generated solutions · Accuracy of AI solutions

1 Introduction

In recent years, artificial intelligence (AI) has significantly transformed the way mathematics is taught and learned at all levels of education, including higher education, particularly engineering studies. AI not only reshapes knowledge transfer but also influences students' analytical skills and challenges educators to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 282–288, 2025.
https://doi.org/10.1007/978-3-031-99264-3_35

adapt to modern teaching methodologies [3]. As AI-powered tools become more prevalent, both students and lecturers are increasingly relying on these technologies to enhance the learning process. One such tool is ChatGPT, a language model capable of generating step-by-step solutions to mathematical problems. While its ability to solve complex equations and explain mathematical concepts is impressive, the accuracy of its responses is not infallible. Mistakes can occur, and it is crucial that students develop critical thinking skills to be able to identify these mistakes.

For educators, integrating AI into mathematics instruction requires not only strong mathematical and pedagogical expertise but also familiarity with AI systems. The evolving role of AI necessitates adjustments in teaching strategies, ensuring that these tools complement rather than replace traditional educational methods [8]. Advances in machine learning algorithms and deep neural networks have made AI more proficient at solving mathematical problems, improving both instruction and research in the field. In engineering courses, for example, AI applications may facilitate the visualization of complex mathematical concepts, making abstract ideas more tangible and practical.

Despite these advantages, AI remains a supportive tool rather than a replacement for human expertise. The ultimate responsibility for effective teaching still lies with educators, who must carefully select instructional methods, content, and assessment strategies [4]. As AI continues to evolve, it is essential to evaluate its accuracy and reliability in solving mathematical problems. This study aims to evaluate the performance of chatbots in solving high school and university-level mathematics problems by analyzing whether students can identify incorrect or flawed solutions. By examining students' ability to critically assess AI-generated responses, we hope to gain insights into both the strengths and limitations of AI in mathematics education.

2 Motivation, Methodology, and Data Collection

In the first semester of the 2023/2024 academic year, a survey was conducted among 636 first-year engineering students at Gdańsk University of Technology. The survey, administered at the beginning and end of the semester, examined students' use of AI tools for academic support. Students were asked, among other things, whether they used any AI tools to support their learning and, if so, which tool they used most often. Students' responses are summarized in Table 1.

The results of the survey indicate a growing usage of AI tools. At the start of the semester, 23% of students reported using AI, which increased to 37% by the semester's end. ChatGPT was overwhelmingly the dominant choice, used by 89% of AI users initially and 92% by the end.

A notable trend was the shift in AI usage patterns: 24% of students who initially did not use AI adopted ChatGPT by the semester's end, whereas 8% stopped using AI tools.

These findings highlight a rising reliance on AI in learning, with ChatGPT being the preferred tool.

Table 1. Results of a survey conducted in the 2023/24 academic year

Question: Do you use any AI tools to support your studies ?		
	start of the semester	end of the semester
Yes	23%	37%
No	77%	63%
Question: If you use any AI tool, which one do you use most often ?		
	start of the semester	end of the semester
ChatGPT	89%	92%
Other	11%	8%

The choice of AI chatbots for the study presented in this paper was guided by their popularity as well as diverse problem-solving approaches. We selected ChatGPT, Perplexity, and Wolfram as they represent different AI architectures and methodologies.

2.1 High School Level

In Poland, since 1999, the system of external primary and high school examinations has been unified across the country. The Central Examination Board (CEB) oversees this system at the K-12 level and is responsible for maintaining the reliability and fairness of external examinations [1].

One of the most significant assessments it oversees is the mathematics Matura exam, held annually as part of the university entrance process. In 2024, the extended-level mathematics paper, based on requirements set by the Minister of Education and Science, included 14 open-ended tasks, with Task 14 selected for this study.

Building on research into artificial intelligence's ability to solve mathematical problems [7], this study examines ChatGPT's performance in solving the chosen exam task and analyzes the types of errors it makes. The research is guided by the following questions:

1. What benefits does ChatGPT offer in teaching, learning, and assessing mathematics?
2. What challenges does ChatGPT face in correctly solving mathematical problems?

The study is based on two hypotheses: (1) ChatGPT can provide personalized support to students, (2) It can adapt to a student's individual learning pace, allowing for flexible understanding of concepts over time.

2.2 University Level

The study, conducted in the 2024/25 academic year, involved 230 s-semester engineering students, all of whom had taken a mid-term mathematics exam a

month earlier. It was conducted on the university's educational platform, where students were given access to a module containing three tasks, each included one example solved by three different AI chatbots: ChatGPT, Perplexity, and Wolfram. In this article, one of the three tasks is chosen for analysis. The data from the remaining two will be analysed in a future, broader study.

Students were asked to:

- Assume the role of a teacher and evaluate the example for its correctness.
- Assess whether evaluating the solution was a difficult task.
- Determine the usefulness of the solution method for self-study.

Additionally, students could provide open-ended feedback on the solutions in a designated comment field. They had two days to complete the task, allowing ample time for thoughtful responses.

The examples analyzed in the study covered material that served as a review for their current coursework. Alongside student evaluations, two academic teachers also assessed the examples, though students were unaware of these grades. The teachers' evaluations were later used in the analysis of the research results.

3 Analysis

3.1 High School Level

In the presented study, we employed the diagnostic survey method, incorporating document analysis, interviews, and questionnaires.

We analyzed a problem from the high school final examination (Matura). The solution proposed by ChatGPT 4.5 was lengthy, complex, and challenging for students to comprehend due to extensive detail. It relied heavily on analytic geometry, an approach rarely applied in practice to this type of problem (see Fig. 1). Despite its complexity, ChatGPT's solution was correct. However, ChatGPT incorrectly labeled the intersection points of the given lines and carried this mistake consistently throughout its calculations.

The solution generated by Wolfram contained a conceptual error, incorrectly identifying the area CEFG as a triangle (see Fig. 1). This misinterpretation led to an incorrect final solution.

3.2 University Level

The task that students were asked to evaluate was: *Using a definite integral, determine the area of the region bounded by the curves* $y = \arccos x, y = \frac{\pi}{3}$ *and* $x = 0$.

The chatbots were given only the task, without any additional prompts. All three began by determining the point of intersection of the curves and setting up the integral to calculate the area of the region. ChatGPT made an error in formulating the integral but still provided a correct numerical answer—without acknowledging the mistake. Of the three, only Perplexity attempted to evaluate

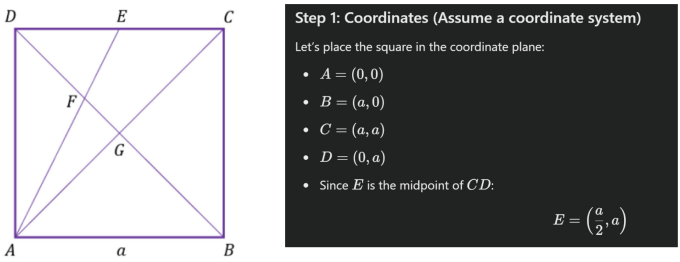


Fig. 1. Illustration for the task and a fragment of the proposed solution

the integral analytically and show the solution steps. ChatGPT and Wolfram, by contrast, provided only final answers: Wolfram gave an exact solution, while ChatGPT offered a numerical approximation. A significant number of students commented on the lack of step-by-step solutions or exact answers, yet still gave the chatbots high assessment marks.

The study data were analyzed using a Colab notebook in Python 3.11.11. The teacher's assessment served as the reference point for comparison. After examining correlations between variables, no strong linear correlation was found, indicating that student evaluations and teacher assessments did not follow a predictable linear pattern. The average difference between student and teacher assessments was computed, and the results were also analyzed using a Wilcoxon signed-rank test. The null hypothesis stated that there were no differences between student and teacher assessments, while the alternative hypothesis stated that students rated chatbot-generated solutions higher than teachers did.

Table 2. Average difference between student and teacher assessment, and $p-$values produced by Wilcoxon test.

Chatbot	Average difference	$p-$values
ChatGPT	1.69	9.72×10^{-37}
Perplexity	0.79	1.14×10^{-24}
Wolfram	1.94	1.24×10^{-37}

The results shown in Table 2 demonstrate that students consistently graded AI-generated solutions higher than the reference assessments provided by the teachers. Specifically, chatbots solutions all showed significant differences in favor of the students' assessments, with p-values far below the conventional threshold for statistical significance. These findings indicate that students may have difficulty identifying errors or deficiencies in chatbot-generated solutions, possibly due to their trust in AI responses or a lack of critical analysis when engaging with these tools.

The consistently high evaluations could also suggest that students view these AI systems as authoritative or accurate problem solvers, which may highlight a potential gap in their ability to critically assess and validate AI-driven solutions.

4 Conclusions

The change in the role of the teacher in the AI era is inevitable [2]. In engineering studies, where mathematics is the foundation of many specialized subjects, teachers need to acquire skills to integrate AI tools in a way that strengthens the understanding of mathematical concepts and their practical applications [5]. Required key competencies include: the ability to assess the quality of solutions generated by AI, designing tasks that require critical thinking, not just simple AI applications, and creating didactic scenarios that use AI as a tool to support, not replace, the thought process.

In the AI-driven era, students have moved from copying information from Wikipedia to using generative AI tools. This shift brings both opportunities and threats to the development of their mathematical and critical thinking skills. A key skill that students need to develop is the ability to work with AI while questioning its results and adding their own insights. In engineering, where solving mathematical problems requires creative approaches and deep conceptual understanding, the ability to critically evaluate AI-generated solutions is essential. At the heart of critical thinking is the synergy between AI capabilities and human creativity, ethical reasoning, and understanding of context. Students must learn to verify the correctness of AI-generated mathematical solutions, understand the methodology and steps to a solution, not just the final result, and identify situations where AI may fail or provide incomplete answers.

The scientific literature presents both positive and negative effects of using AI in mathematics education. Positive aspects include increased access to mathematics education, improving teaching efficiency, and supporting students of different levels of proficiency. However, there is also a risk of technology dependency, which can limit the development of analytical thinking and independent problem-solving skills in students [6].

Future versions of AI models are expected to be even better at dealing with advanced mathematical topics such as group theory, functional analysis, and topology - areas of particular importance in engineering. The future of AI mathematics education will likely include increasing integration of advanced AI tools into the teaching process, but the success of this integration will depend on our ability to use technology in a way that enhances, rather than replaces, human mathematical abilities [9].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. The central examination board. https://cke.gov.pl/en/. Accessed 31 Mar 2025

2. Alexandrowicz, V.: Artificial intelligence integration in teacher education: navigating benefits, challenges, and transformative pedagogy. J. Educ. Learn. **13**(6), 346–364 (2024)
3. Bin Mohamed, M.Z., Hidayat, R., Binti Suhaizi, N.N., Bin Mahmud, M.K.H., Binti Baharuddin, S.N., et al.: Artificial intelligence in mathematics education: a systematic literature review. Int. Electron. J. Math. Educ. **17**(3), em0694 (2022)
4. Dąbrowicz-Tlałka, A.M., Musielak, M.M.: Mathematical education at the academic level - technological changes, modifications of habits and challenges. In: Proceedings of the XX Ogólnopolska Konferencja Nauczania Matematyki w Uczelniach Technicznych, pp. 5–19. Uniwersytet Morski w Gdyni (2023)
5. Guzik, A., et al.: The big book of online education for academics and other teaching professionals (2024)
6. Li, M., Noori, A.Q.: Exploring the nexus of attitude, contextual factors, and AI utilization intentions: a PLS-SEM analysis among primary mathematics teachers in china. Asian J. Math. Educ. **3**(3), 289–311 (2024)
7. Ravšelj, D., et al.: Higher education students' perceptions of chatgpt: a global study of early reactions. PLoS ONE **20** (2025). https://doi.org/10.1371/journal.pone.0315011
8. Wijaya, T.T., Yu, Q., Cao, Y., He, Y., Leung, F.K.: Latent profile analysis of AI literacy and trust in mathematics teachers and their relations with AI dependency and 21st-century skills. Behav. Sci. **14**(11), 1008 (2024)
9. Yi, L., Liu, D., Jiang, T., Xian, Y.: The effectiveness of AI on k-12 students' mathematics learning: a systematic review and meta-analysis. Int. J. Sci. Math. Educ. 1–22 (2024)

Making Generative AI Hallucinations Useful by Reassessing the Troublemaker Agent Strategy

Emmanuel G. Blanchard[1,3(✉)], Jean-Christophe Callahan[2], Idir Akretche[1], Naif Asswiel[1], Lucas Schmitt[1], Amine Kessemtini[1], and Imran S. A. Khan[1]

[1] Laboratoire d'Informatique de l'Université du Mans, Le Mans, France
emmanuel.blanchard@univ-lemans.fr
[2] Centre Hospitalier du Mans, Le Mans, France
jccallahan@ch-lemans.fr
[3] McGill Department of Surgery, Montreal, Canada

Abstract. Hallucinations are a known barrier in using generative AI in education. This paper proposes to use the troublemaker agent strategy in order to turn these flaws into productive assets. We present the Troublemaker Agent Platform (TrAP), which supports this approach and includes a low-cost, eco-friendly architecture based on RAG and SLM for independence from commercial LLMs.

Keywords: Troublemaker Agent · Cognitive Dissonance · Generative AI · Language Models · Retrieval Augmented Generation · Medical Education

1 Introduction

A strong healthcare system requires enough well-trained professionals, but many countries have faced shortages in recent decades, making it hard to meet growing healthcare needs and workforce turnover. While some issues have been addressed, such as France lifting a law limiting medical training enrollment in 2019 [29], the problem persists. The number of new candidates is now constrained by a lack of experts to train them while also handling their clinical duties. Educational technologies can enhance such training efforts [5], and in recent years, Generative AI (GenAI) has made a significant impact in this field, despite several limitations. The most notable of these is its tendency to generate false information, commonly referred to as hallucinations [18].

Our research examines how this flaw could be turned into an advantage using the Troublemaker Agent strategy, a techno-pedagogical approach which was developed in the 1990s [1] and is based on Festinger's cognitive dissonance theory [13]. It involves presenting learners with contradictory statements to encourage critical thinking. We propose leveraging GenAI hallucinations to facilitate the creation of "learning-by-troublemaking" activities that are relevant for medical training and other domains.

Section 2 provides a brief literature review on the use of generative AI in education, and on learning-by-troublemaking. Section 3 describes a study conducted to assess

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 289–296, 2025.
https://doi.org/10.1007/978-3-031-99264-3_36

the feasibility of using generative AI to support learning-by-troublemaking. Section 4 introduces the Troublemaker Agent Platform (TrAP), developed based on insights from the study, and Sect. 5 evaluates variants of its Statement Generation Module.

2 State of the Art

Generative AI and Learning. Generative AI, particularly Large Language Models (LLMs), has recently gained attention for its potential in education. Research shows that LLM-based chatbots can evaluate short answers [27], serve as virtual partners [23], generate educational questions [6], and help design curricula [31]. However, integrating generative AI into education brings challenges [19], including concerns about reliability and transparency [33], and cultural biases [4]. A major concern is "hallucinations"—false information generated by the AI that can mislead users [17, 18]. Retrieval-Augmented Generation (RAG) is a common solution to address this issue [20]. It augments prompts by adding relevant context from a knowledge base (KB), which reduces hallucinations [3] and provides better control over the AI's output [21, 22, 32]. This makes RAG increasingly popular in the AIED field.

Learning-by-Troublemaking. Festinger defines cognitive dissonance as the mental discomfort that arises when a person's beliefs, attitudes, or behaviors are in conflict [13]. In this paper, we use the term *learning-by-troublemaking* to refer to teaching methods that apply cognitive dissonance theory to help learners acquire knowledge or skills. This approach has been used in various educational areas, such as language learning [25], chemistry [11], diversity awareness [24], and promoting e-learning among teachers [9]. In the late 90s, AIED researchers introduced the concept of a Troublemaker Agent [1], a virtual assistant designed to purposely create challenges or dissonance in learners to engage their critical thinking. While this techno-pedagogical strategy showed promise, it faced technological hurdles that limited its use.

Each of these approaches helps address some of the limitations of the other. Hence, we believe their combination could lead to new and effective learning technologies.

3 Generative AI to Facilitate Learning-by-Troublemaking

In learning-by-troublemaking, realism is essential for pedagogical effectiveness, as overly false statements won't create the necessary cognitive dissonance. Therefore, the first research hypothesis we evaluate is:

RH1. A non-specialized LLM can generate false-but-realistic statements to support the creation of learning-by-troublemaking activities?

In order to test RH1, we used ChatGPT-4o (November 2024) [8] and, after a few prompt engineering efforts, we agreed on the following prompt:

```
"Using the question I'm about to give you, generate 3
statements that are: a) distinct, and b) false but realis-
tic for an expert rehabilitation physician. Provide each
statement followed by its explanation, labelled 'expl'."
```

For example, for the question "*What are the proven methods to prevent pneumonia acquired under mechanical ventilation when used as first-line treatment*?", ChatGPT-4o generated 3 statements, including the following one:

> "Routine administration of corticosteroids to reduce pulmonary inflammation is an effective method of preventing ventilator-associated pneumonitis."

According to the medical expert of our team, this is a successful generation of a false-but-realistic statement.

Method. The medical expert of our team identified 11 medical questions, resulting in 33 statements generated using the previous prompt. A panel of medical experts evaluated the **falsehood level** of these statement by rating each of them with the scale: *always false, generally false, generally true, always true*. They could also choose not to answer if outside of their expertise. Then, they assessed the **troublemaking potential** of each statement, i.e., how likely it was to create doubt in two user types—medical externs and residents—using a 5-point Likert scale (*never, rarely, sometimes, often, always*). The panel also answered open-ended questions to further assess the quality of generated statements and the relevance of learning-by-troublemaking in medical education.

Expert Panel Description. The evaluation took place over two weeks at the end of 2024 and beginning of 2025, with 14 medical experts (12 men, 2 women) from a French hospital. Thirteen of them work in intensive care, and one is a cardiologist. The average experience as senior doctors is 14.4 years (SD: 12.7). All participants have experience supervising externs and residents. The significant number of experts who joined our study despite their busy schedules illustrates a great interest for our initiative.

Results and Discussion. The medical experts first evaluated ChatGPT-4o's ability to generate false statements. 100% of experts agreed that 20 out of 33 statements were false, with 75–99% confirming 11 more statements as false. The falsity of the last two statements was more debatable, with 50–74% agreeing on it. Then, the troublemaking potential of the statements was assessed. The panel responses were categorized as *very likely, likely, unlikely, or very unlikely* to be troublemaking. Statements with unclear trends were not categorized. The troublemaking potential was rated higher for externs (1 very likely, 8 likely, 4 unlikely, 0 very unlikely) than for residents (0 very likely, 3 likely, 12 unlikely, 1 very unlikely).

The expert panel responses led us to validates RH1: a non-specialized LLM can generate false-but-realistic statements that are appropriate for learning-by-troublemaking activities. Additionally, experts suggested that the troublemaking potential can frequently be significantly increased with minor adjustments, such as changes in grammar or vocabulary, and highlighted the importance of a debriefing session to avoid misinformation being internalized. Overall, many experts expressed a strong interest in using this approach to offer original and engaging activities to trainees. Finally, while this evaluation focused on generating false-but-realistic statements, it is important to note that learning-by-troublemaking also requires true statements. If all statements are false, students may recognize the pattern and bypass the intended learning objectives [2].

4 The Troublemaker Agent Platform

The results of our previous evaluation confirmed the interest of developing the **Troublemaker Agent Platform** (TrAP), a learning-by-troublemaking environment powered by GenAI. Further reflections led us to the following goals for the system:

– To reduce the time and effort needed for an expert to create, manage, and debrief learning-by-troublemaking activities, while ensuring that they have ways to control and alter these tasks if they want to.
– To create a GenAI-based module to support quality statement generation, that is low-cost, eco-friendly, and protects data privacy.

Figure 1 presents a simplified modular architecture of TrAP.

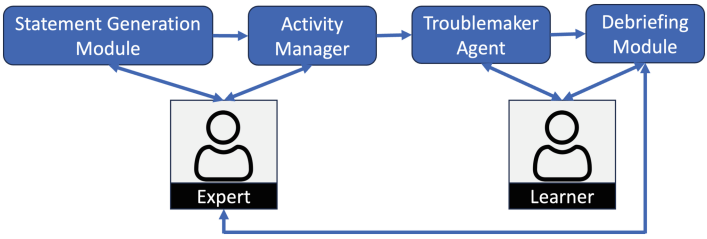


Fig. 1. A simplified modular architecture of TrAP

The **Statement Generation Module** supports experts in creating relevant statements for learning-by-troublemaking. The **Activity Manager** stores previously created statements. It can also support the edition of existing ones or the creation of new statements without AI assistance. Finally, it is used by the expert to group several statements together to create a learning activity. The **Troublemaker Agent** (Fig. 2) then presents this activity to the learner through a simple interface that shows a medical statement

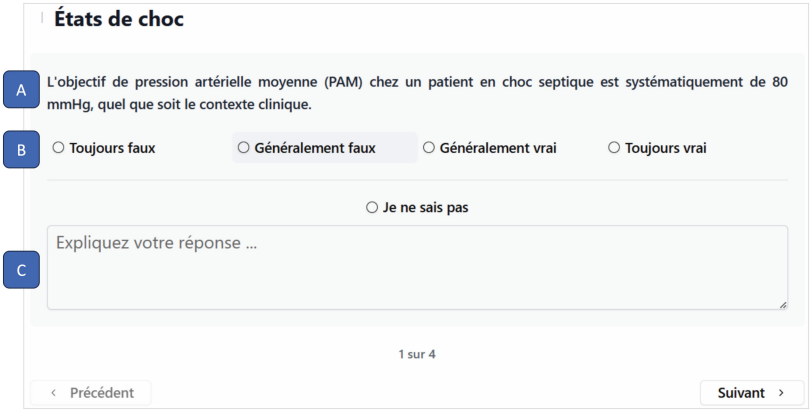


Fig. 2. The graphical interface of TrAP Troublemaker Agent

(A), asks the learner to judge its veracity (B), and to explain their reasoning (C). Logs of user interactions and responses are later used by the **Debriefing Module** to support a feedback session between the learner and the expert.

5 Investigating Statement Generation Module Variants

As mentioned earlier, one of TrAP's goals is to create a low-cost, eco-friendly, and privacy-preserving Statement Generation Module. The aim is to reduce reliance on commercial LLMs like ChatGPT-4o while maintaining good generation quality to assist experts. This is motivated by both ethical and practical concerns, as the hospitals we collaborate with have a limited budget and regulations to protect their data.

With this in mind, we developed two GenAI architecture variants for the Statement Generation Module. The first one draws inspiration from the method used in our preliminary study and relies on API calls to a commercial LLM whereas the second one combines a Small Language Model (SLM) with a RAG process.

The SLM definition we adopt is that of a language model small enough to be hosted on a local machine, which aligns with our goals of being low-cost, eco-friendly, and privacy-preserving. By definition, an SLM has a weaker overall performance than an LLM. However, adding a RAG process will improve the prompt with relevant information retrieved from a KB previously converted into numerical vectors called embeddings (see [20] for the detailed process description). This leads us to the following research hypothesis:

RH2. A statement generation module based on RAG and an SLM can achieve a performance similar to that of a non-specialized LLM.

Model Selection. To evaluate RH2, we chose ChatGPT-4o as our reference LLM, as it was able to generate false but realistic statements in our preliminary study. We also needed to identify promising combinations of embedding models (a key part of the RAG process) and SLMs. The main selection criterion was French language compatibility, since the KB, prompts, and expected outputs are all in French. Based on existing benchmarks, we selected four candidate SLMs [12, 16, 26, 28] and two embedding models [14, 30]. We tested the eight possible combinations using the BERTScore F1 metric [34], which is commonly used to evaluate RAG systems as it measures the semantic fidelity of generated outputs while being robust to reformulations and paraphrases. *Mistral 7B* [26] achieved the highest BERTScore F1 with the *Snowflake* [30] embedding model (0.7575) and also performed well with *French-Me5* [14] embedding model (0.7233). As a result, we selected both RAG configurations for our evaluation, which results in 3 generation conditions: ChatGPT, *RAG Mistral + Snowflake* (hereinafter Snowflake), and *RAG Mistral + French-Me5* (hereinafter Me5).

Statement Generation. Our reference KB consists of 62 PDF documents on Intensive Care Medicine education [7]. The medical expert of our team prepared 3 questions whose answers are found in the KB. For each question, we generated 3 statements with each of the 3 generation conditions, for a total of 27 generated statements.

Method and Panel Description. We gathered a new panel of medical experts to assess the quality of the statements. In addition to evaluating veracity and troublemaking potential, as in Sect. 2, they rated the French quality of the statements on a 5-point Likert scale (*very bad, bad, average, good, very good*) and the effort required to make them presentable and sufficiently troublemaking for learners (*very limited, limited, moderate, high, very high*). Eight medical experts participated in this second round of evaluation. They rated the 27 randomly presented statements based on five criteria: veracity, troublemaking potential for externs, troublemaking potential for residents, French quality, and effort needed. We used Friedman's non-parametric test [15] to analyze the differences in panel judgments for the three conditions, and where significant effects were found, we applied Conover's post hoc comparisons [10] for pairwise analysis.

Results. The Friedman test comparisons revealed that none of the three conditions led to different judgments from our expert panel regarding the troublemaking potential of the generated statements. The ChatGPT and Snowflake conditions were considered very similar overall, with ChatGPT only requiring slightly less effort than Snowflake ($p = .028$). However, the Me5 condition was consistently rated lower, emphasizing the importance of carefully selecting the embedding model. We then measured the retrieval performance of Snowflake using recall@k and precision@k metrics. k is a RAG hyperparameter that determines how many documents are retrieved per request, which affects the retrieval quality. A high precision@k score means the RAG retrieves mostly accurate information, while a high recall@k score means it retrieves as much relevant information as possible, even if some noise is included. For our case, we found $k = 3$ to be optimal, as we prioritized recall@k and noise wasn't an issue for generating false-but-realistic statements. For Snowflake, precision@3 was 0.3137, while recall@3 was 0.9412, i.e. when initialized with $k = 3$, Snowflake retrieved most of the relevant information from the KB related to the medical prompt, while maintaining a limited level of veracity, which is good when generating false-but-realistic statements.

Eventually, the panel's judgment of Snowflake only differing from ChatGPT in terms of required effort led us to accept HR2 with minor reservations: a statement generation module based on RAG and an SLM can achieve an almost equivalent performance to that of a non-specialized LLM.

6 Discussion and Conclusion

In this paper, we validated that generative AI can contribute to the implementation of learning-by-troublemaking activities by providing creative support to experts in a medical context. Learning-by-troublemaking can make constructive use of false-but-realistic statements. Consequently, the hallucinations inherent to generative AI are no longer an issue as long as they are properly monitored (i.e., experts are kept in the loop).

We introduced the Troublemaker Agent Platform (TrAP), which facilitates the creation and management of such activities. To ensure financial and ecological sustainability, we developed statement generative module variants based on RAG and SLM. An evaluation of such modules highlighted that, in the context of supporting learning-by-troublemaking, it is possible to achieve performance close to that of a commercial LLM while being significantly more resource-efficient and ensuring data sovereignty.

In learning-by-troublemaking, debriefing sessions are essential to prevent the internalization of false information. While this is already standard practice in medical education and consequently not a major concern according to our medical expert, it remains important to keep this in mind, as other fields may not be as familiar with this approach.

Future works will focus on enriching debriefing sessions with data visualizations, on running large-scale evaluations of TrAP, and on designing a generator module for true-but-troublemaking statements.

References

1. Aimeur, E., Frasson, C., Dufort, H.: Cooperative learning strategies for intelligent tutoring systems. Appl. Artif. Intell. **14**, 465–489 (2000). https://doi.org/10.1080/088395100403388
2. Baker, R.S.J.D., et al.: Adapting to when students game an intelligent tutoring system. In: Proceedings of ITS 2006. LNCS, vol. 4053, pp. 392–401. Springer (2006). https://doi.org/10.1007/11774303_39
3. Béchard, P., Ayala, O.M.: Reducing hallucination in structured outputs via Retrieval-Augmented Generation. arXiv preprint (2024). arXiv:2404.08189
4. Blanchard, E.G., Mohammed, P.: On cultural intelligence in LLM-based chatbots: implications for artificial intelligence in education. In: Proceedings of AIED 2024. LNCS, vol. 14829, pp. 439–453. Springer (2024). https://doi.org/10.1007/978-3-031-64302-6_31
5. Bridges, S., Chan, L.K., Hmelo-Silver, C.E.: Educational technologies in medical and health sciences education (2015). https://doi.org/10.1007/978-3-319-08275-2
6. Bulathwela, S., Muse, H., Yılmaz, E.: Scalable educational question generation with pre-trained language models. In: Artificial Intelligence in Education, pp. 327–339 (2023)
7. CEMIR. Référentiel du Collège des enseignants de médecine intensive, 2ème cycle. https://www.ce-mir.fr/enseignements/dfasm-edn/livre-referenciel-deuxieme-cycle-r2c. Accessed 4 Apr 2025
8. ChatGPT. https://chatgpt.com/. Accessed 4 Apr 2025
9. Chiou, W.B.: Using cognitive dissonance to enhance faculty members' attitudes toward teaching online courses. Psychol. Rep. **99**(2), 465–471 (2006)
10. Conover, W.J.: Practical nonparametric Statistics, 3rd edn. Wiley (1999)
11. Corradi, D., Clarebout, G., Elen, J.: Cognitive dissonance as an instructional tool for understanding chemical representations. J. Sci. Educ. Technol. **24**, 684–695 (2015). https://doi.org/10.1007/s10956-015-9557-5
12. Deepseek-R1-Distill-Qwen-7B. https://ollama.com/library/deepseek-r1. Accessed 13 Feb 2025
13. Festinger, L.: The arousal and reduction of dissonance in social contexts. In: Schachter, S., Gazzaniga, M. (eds.) Extending Psychological Frontiers: Selected Works on Leon Festinger, pp. 238–257. Russell Sage Foundations, New York (1989)
14. French-Me5-Large. https://huggingface.co/antoinelouis/french-me5-large. Accessed 13 Feb 2025
15. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. **32**(200), 675–701 (1937)
16. Gemma 7B. https://ollama.com/library/gemma. Accessed 13 Feb 2025
17. Hicks, M.T., Humphries, J., Slater, J.: ChatGPT is bullshit. Ethics Inf. Technol. **26** (2024). https://doi.org/10.1007/s10676-024-09775-5
18. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)

19. Kooli, C.: Chatbots in education and research: a critical examination of ethical implications and solutions. Sustainability **15**(7), 5614 (2023)
20. Lewis, P.S.H., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Neural Inf. Process. Syst. **33**, 9459–9474 (2020)
21. Li, X., Henriksson, A., Duneld, M., Nouri, J., Wu, Y.: Supporting teaching-to-the-curriculum by linking diagnostic tests to curriculum goals: using textbook content as context for retrieval-augmented generation with large language models. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. AIED 2024, vol. 14829. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64302-6_9
22. Liu, C., Hoang, L., Stolman, A., Wu, B.: HiTA: a RAG-based educational platform that centers educators in the instructional loop. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. AIED 2024, vol. 14830. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64299-9_37
23. Ma, Q.C., Wu, S.T., Koedinger, K.: Is AI the better programming partner? Human-human pair programming vs human-AI pAIr programming. In: Proceedings of the AIED Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation, pp. 64–77 (2023)
24. McFalls, E.L., Cobb-Roberts, D.: Reducing resistance to diversity through cognitive dissonance instruction: implications for teacher education. J. Teach. Educ. **52**(2), 164–172 (2001). https://doi.org/10.1177/0022487101052002007
25. Miao, Y., Moran, M., Kang, O.: A cognitive dissonance approach to moderating listener perception of L2 English speakers. J. Multilingual Multicultural Dev. 1–18 (2023). https://doi.org/10.1080/01434632.2023.2236073
26. Mistral 7B: https://ollama.com/library/mistral. Accessed 13 Feb 2025
27. Nye, B.D., Mee, D., Core, M.G.: Generative large language models for dialog-based tutoring: an early consideration of opportunities and concerns. In: Proceedings of the AIED Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation, pp. 78–88 (2023)
28. Phi3. https://ollama.com/library/phi3. Accessed 13 Feb 2025
29. Roberto, A.: La fin du numérus clausus n'aura pas d'effet avant 2035. Fondation pour la recherche sur les administrations et les politiques publiques. https://www.ifrap.org/emploi-et-politiques-sociales/la-fin-du-numerus-clausus-naura-pas-deffets-avant-2035. Accessed 13 Feb 2025
30. Snowflake-Arctic-Embed2: https://ollama.com/library/snowflake-arctic-embed2. Accessed 13 Feb 2025
31. Sridhar, P., Doyle, A., Agarwal, A. Bogart, C., Savelka, J., Sakr, M.: Harnessing LLMs in curricular design. Using GPT-4 to support authoring of learning objectives. In: Proceedings of the AIED Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation, pp. 139–150 (2023)
32. Thüs, D., Malone, S., Brünken, R.: Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature. Front. Psychol. **15**, 1474892 (2024). https://doi.org/10.3389/fpsyg.2024.1474892
33. Wu, Y., Henriksson, A., Duneld, M., Nouri, J.: Toward improving the reliability and transparency of ChatGPT for educational question answering. In: Proceedings of EC-TEL2023. LNCS, vol. 14200, pp. 475–488. Springer (2023)
34. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: Proceedings of ICLR2020 (2020). http://arxiv.org/abs/1904.09675. https://doi.org/10.48550/arXiv.1904.09675

Enhancing Collaborative Academic Writing with Generative Artificial Intelligence: Use and Effects

Anna Korchak[1(✉)], Mik Fanguy[2], Kseniia Adamovich[1], Han Zhang[1], Mattew Baldwin[3], and Jamie Costley[4]

[1] National Research University Higher School of Economics, 20 Myasnitskaya Ulitsa, 101000 Moscow, Russia
aekorchak@hse.ru
[2] The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong, China
[3] Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Guseong-dong, Yuseong-gu, Daejeon, South Korea
[4] United Arab Emirates University, H1 Sheik Khalifa Bin Zayed Street, Al Ain, United Arab Emirates

Abstract. The increasing use of Generative Artificial Intelligence (Gen-AI) in higher education offers a lot of opportunities to enhance collaborative writing. However, educators face challenges in understanding how students use these tools in group settings, which complicates their effective integration into the classroom. Moreover, little is known about the effects of specific Gen-AI applications. To fill this gap, the present study aims to identify particular Gen-AI applications in group writing tasks and examine how they relate to students' group performance. The study focuses on analyzing how students use Gen-AI in group work, particularly when completing writing tasks. The sample includes 143 Korean students enrolled in an academic writing course, who were asked to collaboratively write a section of an academic paper. Based on students' written self-reflections and identified Gen-AI applications in their texts, three categories of Gen-AI use were elicited: conceptual, language, and integrated. Conceptual applications involved tasks such as brainstorming, summarizing, paraphrasing, and other content-related activities. Language applications focused on grammar correction, vocabulary enhancement, and similar linguistic tasks. Integrated applications combined both conceptual and language uses. Conceptual and integrated applications were found to positively impact group performance, while language applications did not. These findings contribute to the broader discourse on Gen-AI use in education, with a particular focus on group-based applications and their effects.

Keywords: Academic Writing · Artificial Intelligence · Collaborative Writing · Generative Artificial Intelligence · Student Performance

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 297–304, 2025.
https://doi.org/10.1007/978-3-031-99264-3_37

1 Introduction

Gen-AI tools have gained popularity in higher education, particularly among university students who often engage in writing tasks [1]. The widespread use of these tools is driven by the benefits they offer, including enhanced academic performance, cognitive development, and improved creativity [2, 3]. While a number of studies have investigated how Gen-AI is applied in individual writing tasks [4–6], research on its use in group work remains scarce. Although some works highlight the significant role of Gen-AI in collaborative contexts [7, 8], they briefly touch upon the topic and do not provide details of how Gen-AI is used and the exact effects. As a result, it is unclear whether Gen-AI use in individual and group settings is different. Understanding specific Gen-AI uses in group work is particularly crucial, as Gen-AI's interactive nature allows it to act as a potential collaborator [9]. Considering the benefits that Gen-AI brings, such as interactivity, engagement, and improved performance [2, 3], these can be reinforced by the advantages of collaborative learning. The latter include students achieving better results when working together and completing tasks more efficiently [10, 11]. Therefore, the benefits of collaborative learning and Gen-AI use may be reinforced when combined—if Gen-AI is treated as a real peer. However, this requires a clear understanding of how students actually apply Gen-AI in group work.When it comes to the effects of Gen-AI use on learning, existing studies have primarily focused on individual applications. It is suggested that Gen-AI facilitates students' thinking and help them quickly grasp new topics [12, 13]. Supports students in becoming better writers and improves their performance [14]. The outcomes of Gen-AI use in group settings remain less understood. While some research highlights Gen-AI's potential to enhance group interactions and idea exchange [7], other studies report minimal impact on group performance [8]. This inconsistency highlights the need for a closer examination of specific Gen-AI applications in group contexts to determine which uses lead to the most beneficial outcomes. Therefore, the present study aims to research particular uses of Gen-AI within group settings and assess the results of these uses in terms of performance. The research question we seek to answer are as follows:

RQ1: How do students specifically use Gen-AI when working in groups to complete writing tasks?
RQ2: What are the effects of these specific Gen-AI uses on students' group performance?

2 Literature Review

Studies examining how Gen-AI is applied in education highlight that its interactive applications tend to involve more conceptual uses, whereas superficial applications are limited to routine tasks [6]. According to [15], the most common reasons students use Gen-AI in academic settings include asking clarifying questions and gaining a deeper understanding of subject-specific concepts. In the context of writing specifically, research indicates that Gen-AI is used for various language-related purposes. For instance, [5] note its application in vocabulary enhancement, [4] highlight its use in improving sentence construction, and [16] emphasize its role in assisting with translations. Beyond language aspects, Gen-AI also supports content-related tasks. It has been found to help students

summarize and paraphrase information for inclusion in their writing [16]. Additionally, Gen-AI is commonly used both in individual work [4, 6] and group collaborations [17] for working with ideas students need for their writing. Feedback generated by AI has been shown to contribute not only to measurable improvements in students' essay revisions but also to positively influence their engagement levels, motivation, and emotional responses [18]. Research indicates that ChatGPT significantly enhances students' academic writing skills when they interact with it on an individual basis [19]. After engaging with generative AI during the writing process, students reported that it was particularly useful for checking grammar, performing translations, and outlining writing tasks [20]. In terms of skill development, studies have found that generative AI helps students improve their writing abilities and boosts their self-efficacy [20]. Even in the short term, using AI-assisted tools for academic writing has been significantly linked to improvements in students' scores [14]. However, some studies suggest that reliance on generative AI may have adverse effects on students' writing abilities [21].

3 Method

3.1 Procedures

Students from a Korean university enrolled in an academic writing course were divided into groups of 4 or 5 and tasked with writing the Discussion of an academic paper based on results provided by the course instructor. Prior to the task, students watched a video lecture on how to write a Discussion. The instructor did not provide any specific training on using Gen-AI but encouraged students to apply it in any form they wanted. The groups collaborated in Google Docs and were given 60 min to complete the task. Students used different font colours to indicate their individual contributions to the group writing task and highlighted any text generated with the help of Gen-AI in yellow. Upon completion, students were asked to submit a written reflection describing in detail how they used Gen-AI during the task. Prior to conducting the research the IRB approval from the university ethical research committee was obtained.

3.2 Participants

The study involved STEM-major students enrolled in a scientific writing course. Of the 154 students, 143 participated in the Introduction task, while 11 opted out. The sample included 81 males (56%) and 61 females (43%), with an average age of 26 years (SD = 3.5). By study level, 83 students (58%) were in master's programs, 16 (11%) in PhD programs, and 44 (30%) in integrated master's-doctoral programs. Before participating in the present study, all students provided informed consent, indicating their agreement to take part in the research and permitting the future anonymous use of their data.

3.3 Assessment

To ensure consistent grading, raters used standardized rubrics adapted from [22]. The rubrics focused on five elements, each rated on a scale from 0 to 2, for a maximum total

of 10 points. First, we examined the use of a reverse funnel structure, in which students transition from specific findings to broader implications. We also assessed whether key results were restated or highlighted clearly and concisely. Another criterion was the inclusion of students' own interpretations or insights, demonstrating their ability to think critically about their findings. Additionally, we evaluated the clarity, organization, and conciseness of the section, ensuring that it was easy to follow and well-structured. Finally, we checked whether students included reflections on the limitations of their work and directions for future research. The tasks were assessed by two raters, with a third resolving disagreements. Scores and feedback were for learning only, while the individual assignment counted for 10% of the course grade.

3.4 Data and Analysis

In order to elicit the exact uses of Gen-AI, thematic analysis [23] was applied. The analysis followed a six-phase process: familiarization with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the final report. Two research team members worked independently to code students' reflections on Gen-AI use and held regular meetings to discuss repeated themes that emerged until the inter-coder reliability reached 90%. To establish a relationship between group and individual scores and the way Gen-AI was utilised, OLS regression analysis was applied. The total word count of each student's contributions to the final group documents is referred to as the volume. The use of different font colours, which identified individual contributions inside the joint work, made this count easier. The AI volume, which is indicated by yellow-highlighted text, is the total number of words produced by Gen-AI in each student's contributions. In this study the dependent variable is the group score on the collaborative writing task, as it represents the outcome we aim to explain. The independent variables include the types of Gen-AI use (language, conceptual, integrated, and unknown), as well as the group mean volume of writing and the group mean volume of Gen-AI-generated text. These variables are examined to assess their potential influence on group performance. Additionally, the pretest score is included as a control variable to account for the initial academic level of each group. By controlling for prior knowledge or ability, the analysis aims to isolate the effects of Gen-AI usage on the writing outcomes more accurately. During the analysis, all students were anonymised—their names were replaced with randomly generated identification numbers created specifically for this study.

4 Results

The thematic analysis of students' reflections revealed three ways Gen-AI was used: language-based, conceptual, and integrated applications. In terms of language-based use, students employed Gen-AI to help with error and grammar correction, improving vocabulary and sentences, translation, information retrieval, and citation formatting. For example, one student explained, "I wrote a draft version of my writing in English and then asked ChatGPT if there were any grammatical mistakes or inappropriate expressions," illustrating the common practice of using AI as a grammar-checking tool. Another

reported, "I examined the whole sentence using AI to check typos and grammar errors," showing how AI was used to polish the language. For conceptual applications, Gen-AI was utilized to support such tasks as summarising, paraphrasing, categorising content, generating or adapting outlines, drafting full texts, and working with ideas. One student described how they "asked GPT to summarise the reference paper that [they] would like to incorporate in [their] paragraph," while another shared that AI helped them "rephrase the original text while maintaining its message in a great expression." Another example highlighted how AI facilitated conceptual organization: "I made AI categorise the summaries of given articles into 3–4 groups based on their applicability for an introduction." Lastly, students reported integrated uses of Gen-AI, where both language and conceptual functions were combined. For example, these uses involved asking AI to help write a paragraph and then using it again to polish the language or check logical flow. The regression analysis revealed several significant predictors of group performance in the collaborative writing task. First, group mean volume of writing was positively associated with group scores ($\beta = 0.009$, $p < 0.01$), indicating that higher overall contributions from group members correlated with better performance. However, the group mean volume of Gen-AI-generated content was not significantly related to group scores ($\beta = 0.002$, $p > 0.1$). When examining specific Gen-AI uses, both conceptual Gen-AI use ($\beta = 0.823$, $p < 0.01$) and integrated use ($\beta = 1.044$, $p < 0.05$) showed statistically significant positive effects on group scores. Interestingly, groups where Gen-AI use could not be clearly categorized also showed a significant positive effect ($\beta = 1.196$, $p < 0.05$). In contrast, language-only Gen-AI use did not significantly impact group performance ($\beta = 0.360$, $p > 0.1$). The overall model accounted for approximately 20% of the variance in group scores (Adjusted $R^2 = 0.158$, $F(7, 136) = 4.829$, $p < 0.01$), supporting the relevance of the identified predictors.

5 Discussion

The present study aimed to explore how students use Gen-AI in collaborative writing tasks and to assess the effects of these uses on group performance. The findings reveal three general applications of Gen-AI: language-based, conceptual, and integrated, and highlight their differential impacts on group outcomes. Conceptual and integrated uses were found to significantly enhance group performance, while language-based applications were not. These results contribute to the growing body of literature on Gen-AI in education, particularly in the context of collaborative learning [6, 15]. By providing detailed insights into how students interact with Gen-AI during group writing tasks, this study addresses a critical gap identified by [17], who emphasized the need for a deeper understanding of Gen-AI's role in collaborative settings. Furthermore, the findings align with earlier research suggesting that AI tools can support higher-order thinking and idea generation when used strategically [16, 20]. For educators, these insights offer practical guidance on integrating Gen-AI into collaborative tasks to maximize its pedagogical value. As for implications of the findings, the positive effects of conceptual and integrated Gen-AI applications underscore the importance of using AI as a tool for stimulating higher-order thinking and idea generation. When students used Gen-AI

to summarize, paraphrase, categorize content, or refine outlines, they were better able to produce well-structured, coherent writing. This aligns with prior research suggesting that AI can facilitate deeper engagement with subject matter and support critical thinking [8]. Integrated uses, which combine both conceptual and language functions, further demonstrate the potential of Gen-AI to act as a versatile collaborator, assisting students in refining both the content and presentation of their ideas. In contrast, the lack of impact from language-based Gen-AI use suggests that while such applications may streamline the writing process by correcting grammar or enhancing vocabulary, they do not necessarily contribute to improved group performance. This finding aligns with concerns raised in previous studies about the over-reliance on AI for surface-level tasks, which may detract from opportunities for deeper learning [21]. For educators, this implies that encouraging students to focus on conceptual and integrated uses of Gen-AI could maximize its pedagogical value, particularly in collaborative settings.

5.1 Limitations and Future Research Directions

The limitations of the study are as follows. First, the data is self-reported, and students' responses may not accurately reflect their actual use of Gen-AI tools. Future research could incorporate more objective measures, such as tracking Gen-AI use logs. Second, the study did not measure students' AI literacy, which is a significant factor for the way Gen-AI is used. Future studies should consider incorporating AI literacy as a variable and assess its impact on student-Gen-AI interaction. Third, while this study focused on the use of Gen-AI in writing tasks, future research could explore how Gen-AI is applied across different types of tasks: presentations and problem-solving activities. Another direction is to examine how Gen-AI impacts long-term learning. The current study focused on short-term group performance, but it would be valuable to explore how use of Gen-AI tools affects students' writing skills development over time.

5.2 Practical Recommendations

Based on the study's findings, there are several recommendations. First, encourage conceptual engagement with Gen-AI: focus on applications as working with ideas, summarization, paraphrasing, and content organization. Second, promote integrated use of Gen-AI across all writing stages via instructing students to use Gen-AI both during the ideation and revision stages. Third, de-emphasize language-only use of Gen-AI, as relying on it solely for these surface-level corrections does not contribute significantly to learning outcomes. Fourth, include Gen-AI use as a visible part of collaboration. Adopt colour coding or version tracking to help students document their use of Gen-AI and ask students to reflect on their Gen-AI use after the task is completed.

Acknowledgments. The authors have no acknowledgements to make.

Data Availability. The dataset used and/or analysed during the current study is available from the corresponding author on reasonable request.

AI Use Declaration. We used generative AI tools (ChatGPT) to improve grammatical aspects of the papers.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Rudolph, J., Tan, S., Tan, S.: ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? J. Appl. Learn. Teach. **6**(1), 342–363 (2023). https://doi.org/10.37074/jalt.2023.6.1.9
2. Essel, H.B., Vlachopoulos, D., Essuman, A.B., Amankwa, J.O.: ChatGPT effects on cognitive skills of undergraduate students: receiving instant responses from AI-based conversational large language models (LLMs). Comput. Educ. Artif. Intell. **6**, 100198 (2024). https://doi.org/10.1016/j.caeai.2023.100198
3. Grájeda, A., Burgos, J., Córdova, P., Sanjinés, A.: Assessing student-perceived impact of using artificial intelligence tools: construction of a synthetic index of application in higher education. Cogent Educ. **11**(1), 2287917 (2024). https://doi.org/10.1080/2331186X.2023.2287917
4. Woo, D.J., Susanto, H., Yeung, C.H., Guo, K., Fung, A.K.Y.: Exploring AI-generated text in student writing: How does AI help? arXiv preprint arXiv:2304.02478 (2023). https://arxiv.org/pdf/2304.02478
5. Ippolito, D., Yuan, A., Coenen, A., Burnam, S.: Creative writing with an AI-powered writing assistant: Perspectives from professional writers. arXiv preprint arXiv:2211.05030 (2022). https://arxiv.org/pdf/2211.05030
6. Nguyen, A., Hong, Y., Dang, B., Huang, X.: Human-AI collaboration patterns in AI-assisted academic writing. Stud. High. Educ. 1–18 (2024). https://doi.org/10.1080/03075079.2024.2323593
7. Zhang, B.: Exploring pedagogical ideologies and strategies for college English writing instruction from a generative artificial intelligence perspective. J. Contemp. Educ. Res. **8**(1), 173–178 (2024). http://ojs.bbwpublisher.com/index.php/JCER/article/download/5951/5129
8. Hu, W., Tian, J., Li, Y.: Enhancing student engagement in online collaborative writing through a generative AI-based conversational agent. Internet High. Educ. 100979 (2024). https://doi.org/10.1016/j.iheduc.2024.100979
9. Kim, J., Yu, S., Detrick, R., Li, N.: Exploring students' perspectives on generative AI-assisted academic writing. Educ. Inf. Technol. 1–36 (2024). https://doi.org/10.1007/s10639-024-12878-7
10. Barkley, E.F., Major, C.H.: Student Engagement Techniques: A Handbook for College Faculty. Wiley (2020)
11. Bielaczyc, K., Collins, A.: Learning communities in classrooms: a reconceptualization of educational practice. In: Instructional-Design Theories and Models, pp. 269–292. Routledge (2013)
12. Salinas-Navarro, D.E., Vilalta-Perdomo, E., Michel-Villarreal, R., Montesinos, L.: Using generative artificial intelligence tools to explain and enhance experiential learning for authentic assessment. Educ. Sci. **14**(1), 83 (2024). https://doi.org/10.3390/educsci14010083
13. Chen, Y., Jensen, S., Albert, L.J., Gupta, S., Lee, T.: Artificial intelligence (AI) student assistants in the classroom: designing chatbots to support student success. Inf. Syst. Front. **25**(1), 161–182 (2023). https://doi.org/10.1007/s10796-022-10291-4
14. Nazari, N., Shabbir, M.S., Setiawan, R.: Application of artificial intelligence powered digital writing assistant in higher education: randomised controlled trial. Heliyon **7**(5) (2021). https://doi.org/10.1016/j.heliyon.2021.e07014

15. Von Garrel, J., Mayer, J.: Artificial intelligence in studies—use of ChatGPT and AI-based tools among students in Germany. Humanit. Soc. Sci. Commun. **10**(1), 1–9 (2023). https://doi.org/10.1057/s41599-023-02304-7
16. Delcker, J., Heil, J., Ifenthaler, D., Seufert, S., Spirgi, L.: First-year students AI-competence as a predictor for intended and de facto use of AI-tools for supporting learning processes in higher education. Int. J. Educ. Technol. High. Educ. **21**(1), 18 (2024). https://doi.org/10.1186/s41239-024-00452-7
17. Buyserie, B.: Collaborative writing and AI: A research assignment for an undergraduate professional communication course. In: Teaching and Generative AI (2024). https://uen.pressbooks.pub/teachingandgenerativeai/chapter/collaborative-writing-and-ai-a-research-assignment-for-an-undergraduate-professional-communication-course/
18. Chan, S., Lo, N., Wong, A.: Leveraging generative AI for enhancing university-level English writing: comparative insights on automated feedback and student engagement. Cogent Educ. **12**(1), 2440182 (2025). https://doi.org/10.1080/2331186X.2024.2440182
19. Mahapatra, S.: Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. Smart Learn. Environ. **11**(1), 9 (2024). https://doi.org/10.1186/s40561-024-00295-9
20. Malik, A.R., Pratiwi, Y., Andajani, K., Numertayasa, I.W., Suharti, S., Darwis, A.: Exploring artificial intelligence in academic essay: higher education student's perspective. Int. J. Educ. Res. Open **5**, 100296 (2023). https://doi.org/10.1016/j.ijedro.2023.100296
21. Niloy, A.C., Akter, S., Sultana, N., Sultana, J., Rahman, S.I.U.: Is ChatGPT a menace for creative writing ability? An experiment. J. Comput. Assist. Learn. **40**(2), 919–930 (2024). https://doi.org/10.1111/jcal.12929
22. Clabough, E.B., Clabough, S.W.: Using rubrics as a scientific writing instructional method in early stage undergraduate neuroscience study. J. Undergrad. Neurosci. Educ. **15**(1), A85 (2016)
23. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)

Exploring Students' Generative AI Usage Patterns and Knowledge Creation in Collaborative Problem Solving

Shotaro Naganuma[1](✉), Tsubasa Minematsu[2], Sachika Shibukawa[3], Aikana Ohno[1], and Yukinori Wakihama[1]

[1] Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, Japan
naganuma.shotaro.062@m.kyushu-u.ac.jp
[2] Kyushu Institute of Technology, 680-4 Kawazu, Iizuka-shi, Fukuoka, Japan
[3] Chuo University, 742-1 Higashinakano, Hachioji-shi, Tokyo, Japan

Abstract. Understanding influence of students' use of generative AI (GenAI) on their learning outcomes is one of the critical issues in the current education sphere, whereas impacts on group-level learning outcomes in creative collaborative problem solving contexts are unexplored. We present one of the first studies to preliminarily investigate how university students' GenAI use in collaborative learning settings related to their knowledge creation quality as group-level learning outcomes, where students worked toward designing future classroom. We found that 1) students used GenAI much more to visually embody their ideas than to find more innovative ideas beyond their own ideas, 2) frequent use of GenAI were linked with middle-level learning outcomes rather than high-level learning outcomes. These findings suggests that while GenAI can support collaborative problem solving moderately, higher learning outcomes require effective integration of human creativity with AI-generated content. Further study is required to identify how high-outcome groups collaboratively developed their ideas using GenAI and what support is effective for students' more constructive use of GenAI for knowledge creation.

Keywords: Generative AI · Computer-Supported Collaborative Learning (CSCL) · Knowledge Creation · Collaborative Problem Solving · Jigsaw Method

1 Introduction

"Knowledge age" [6], in which the creation of new knowledge becomes a necessity, demands education to prepare students to develop the capacity to inquire, innovate, and engage in creative knowledge work and progressive discourse [4, 11, 13]. In response to this challenge, Computer-Supported Collaborative Learning (CSCL) began to develop. CSCL is built on the premise that collaborative knowledge creation and problem solving can effectively be assisted by technology. As such technology, five types were identified [8]: Communication technology, Dynamic technologies, Sharing and co-construction tools, Systems or environments, and Hardware. For this line of study, generative AI

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 305–312, 2025.
https://doi.org/10.1007/978-3-031-99264-3_38

(GenAI), an emerging technology, is one of the promising tools that can support CSCL by incorporating multiple characteristics from those five types.

Use of GenAI exploded from late 2022 among students especially in higher education institutions [1, 3, 14]. Ravšelj et al. found that students primarily used GenAI for brainstorming, summarizing texts, and finding research articles [12]. In the context of CSCL, Chen et al. concluded that GenAI can play a constructive role in group knowledge creation by providing a springboard for discussion when students used it for distinct purposes (e.g., propose hypotheses, suggest revisions) [5]. They urge educators to create scaffolds for "mindfully engaging with GenAI in the classroom" [5] rather than banning it. Several studies, however, theoretically cautioned its negative effects. For example, Dwivedi et al. pointed out that heavy use of ChatGPT might kill creativity and critical thinking [7]. Yilmaz and Yilmaz also indicated that continuous use of GenAI may foster student laziness and weaken their cognitive skill development [17]. To determine whether and how GenAI can be integrated into collaborative learning for knowledge creation, we need to understand how learners utilize GenAI in collaborative learning contexts, to what extent they employ it, and how these uses influence their knowledge creation in collaborative problem solving.

Therefore, we aimed to investigate how university students' use of GenAI in collaborative learning settings related to their knowledge creation as learning outcomes in collaborative problem solving. The following two research question guided our study:

RQ1. How do students use GenAI in a collaborative problem-solving?
RQ2. How does use of GenAI relate to creative artifacts?

To address these, we implemented collaborative learning classroom informed by jigsaw method—one of the collaborative learning models—, where students have open access to GenAI. Students' GenAI use records and their artifacts were used for our analysis.

2 Related Work

Recent studies have explored the integration of AI into jigsaw method. Cao et al. designed a conversational AI partner that facilitated small-group collaboration by generating prompts, encouraging participation, and adapting to group dynamics [2]. Similarly, Lee et al. developed an AI-powered voice assistant participated as a peer in jigsaw activities [9]. They demonstrated that AI use enhanced learners' pedagogical skills. Both studies highlight the potential of AI in collaborative learning, suggesting new directions for human–AI collaboration.

While previous studies have effectively explored task-specific AI designed for particular contexts, our research highlights how learners interact with general-purpose GenAI that are publicly accessible and widely used (e.g., ChatGPT, Dify). Especially, we are interested in group-level learning outcomes rather than individual-level learning gains and in exploring the diverse ways in which learners employ GenAI—from image generation to fact-seeking—as part of their collaboration process.

3 Method

3.1 Context, Participants, and GenAI Implementation

This study was conducted in a required course at a Japanese national university where all 2,500 freshmen participate. Students engage in problem-solving activities with others from various departments. They were divided into several classes and completed 4-week lessons under different instructors. Each week has consecutive two 90-min sessions. Our students were required to design "Future classroom of our university in 2050" as knowledge creation. Our analysis included data from 141 students.

We used Dify—an open-source publicly accessible platform. Our GenAI contains several apps (Google, arXiv, PubMed, dalle, chart, webscraper) and has a ChatGPT-like user interface. Students used natural language prompts to search information, access academic literature, create illustrations, etc. Following Week 1 orientation, students had unlimited access to GenAI.

3.2 Lesson Design

We designed our lesson based on jigsaw method for constructive interaction for innovating ideas [10]. Week 1 served as an orientation where course overview and evaluation criteria were explained. Students were assigned to groups of 4–5 members and introduced to our GenAI. After practicing GenAI, students designed their future classroom illustration using GenAI without expert knowledge. They then were introduced to the four essential components as expert knowledge: space design, artifacts, community, and activity. In Week 2, students participated in expert groups for understanding the four components. They then returned to their original groups to integrate their expert knowledge. Students were tasked with updating their classroom designs with additional information to be researched for Week 3. Weeks 3 and 4 focused on presentation preparation and delivery. Students shared their findings they researched for Week 3, finalized their presentation slides, and delivered 5-min presentations incorporating the four elements. Instructors evaluated these presentations based on innovation and coverage of the essential components (see below for rubric).

3.3 Data Collection

Students' GenAI usage analyzed by referring to the coding framework adapted from Tong and Chan [15], which focused on discourse moves toward knowledge creation. Initially using their relevant codes (e.g., fact-seeking, explanation-seeking), the coding process involved two coders who independently analyzed groups' GenAI usage records in multiple phases. Through iterative discussions and revisions, we developed the final coding framework (Table 1) and completed coding of all GenAI usage records.

Student presentations were evaluated by three professional researchers who developed expert reading materials using a rubric with seven perspectives. It includes "Understanding of expert material (A-D, respectively)", "Component harmony", "Component Expansion", and "Presentation quality." Scores ranged from 0–4 for each perspective (total: 0–28). The rubric, based on SOLO taxonomy, was refined through collaborative

evaluation of initial groups. It was then used for evaluating all 29 groups' presentations. All scores were determined through discussions among the three professors.

Table 1. Final coding framework

Codes	Descriptions
Fact-seeking	Questions seeking factual information. (e.g., *What are learning commons? What is social anxiety?*)
Explanation-seeking	Questions seeking open-ended responses with explanations. (e.g., *What will future classrooms look like? What are the reasons for social anxiety disorder?*)
Sustained inquiry	Asking further questions based on previous ideas and thus deepening the discussion deeper. (e.g., *Can you evaluate our future classroom design? Give us another idea to improve our current design.*)
Simple claim	Simple (dis)agreement or repetition of a statement. (e.g., *Sounds Nice! Good!*)
Illustration request	Request of visual representations or illustrations to aid in understanding or presentation. (e.g., *Create an illustration of an accessible lecture room.*)
Prompt generation request	Request for support in generating or refining effective AI prompts. (e.g., *What prompts can I use when I need ideas?*)
GenAI usage inquiry	Inquiry related to how to use the GenAI tool, including its features, functions, or limitations. (e.g., *How can I effectively communicate my vision for an illustration to AI?*)
Information seeking for group regulation	Inquiry aimed at coordinating or managing group work, such as organizing roles or rules. (e.g., *What rules should I present to my group members?*)
Off-task/practice inquiry	Input unrelated to the main task or used for practice during initial trial sessions. (e.g., *Tell me how to get up early.*)
Unfinished entry	Incomplete input, likely resulting from accidental submission before the message was fully written. (e.g., *Create an illustration of.*)
Unclear	Input lacking sufficient clarity or context to determine its intent or category such as randomly typed alphabets
Others	Input that does not fall under any other categories

3.4 Data Analysis

Concerning on RQ1, we employed a descriptive coding analysis of their GenAI usage records. The frequency of each code was counted and averaged. To investigate the relationship between GenAI usage and group-level learning outcomes (RQ2), we compared the frequency of each GenAI usage by dividing the 29 groups into several categories based on their presentation scores. Presentation scores ranged from 11 to 23 points. The distribution was as follows: 11points (2 groups), 13 (3), 14 (8), 15 (3), 16 (2), 17 (2), 18 (5), 19 (1), 21 (1), 22 (1), 23 (1). Groups were categorized into three performance levels: low (11–13 points, 5 groups), middle (14–18 points, 20 groups), and high (19–23 points, 4 groups). If use of GenAI leads to a more creative artifact, then we can observe a simple linear relationship that higher learning outcome groups use GenAI more frequently. One-way ANOVA with Holm-corrected post-hoc comparisons was used to analyze differences in GenAI usage across these three performance levels.

4 Results

4.1 Frequency of GenAI Usage on Each Code (RQ1)

Table 2 shows average frequency of each GenAI usage code. Out of the four codes from Tong and Chan [15]—Fact-seeking, Explanation-seeking, Sustained inquiry, and Simple claim—, only "Sustained inquiry" was over one-time use in average, with maximum use reached to 18. Out of other four codes—Illustration request, Prompt generation request, GenAI usage inquiry, and Information seeking for group regulation—, "Illustration request" appeared most frequently (M = 29.91), while its variance is large. Among the remaining marginal four codes, "Off-task/practice inquiry" and "Unfinished entry" appeared relatively frequently.

Table 2. Average frequency of each GenAI usage code

Codes	Mean (SD)	Min	Max
Fact-seeking	0.54 (0.95)	0	6
Explanation-seeking	0.39 (1.01)	0	7
Sustained inquiry	1.85 (3.29)	0	18
Simple claim	0.59 (1.79)	0	12
Illustration request	29.91 (23.47)	0	172
Prompt generation request	0.05 (0.28)	0	2
GenAI usage inquiry	0.13 (0.47)	0	3
Information seeking for group regulation	0.28 (1.38)	0	11
Off-task/practice inquiry	13.19 (7.73)	0	48
Unfinished entry	2.09 (3.35)	0	26
Unclear	0.02 (0.14)	0	1
Others	0.29 (1.06)	0	9

4.2 Relationship Between GenAI Usage and Learning Outcomes (RQ2)

We analyzed the GenAI usage records of students in each group (Fig. 1). Out of all the codes analyzed, only two codes—"Illustration request" and "Off-task/practice inquiry"—showed statistically significant differences ($p < 0.05$). For the "Illustration request" code, a significant difference was found among the three groups (F (2, 138) $= 6.013$, $p = .003$). Post-hoc comparisons indicated that the middle-score group used the "Illustration request" entry significantly more than the low-score group (adjusted $p = .007$, $d = .690$). No significant differences were found between the middle- and high-score groups, nor between the high- and low-score groups.

For the "Off-task/practice inquiry" code, a significant difference was found among the three groups (F (2, 138) $= 5.669$, $p = .004$). Post-hoc comparisons indicated that the middle-score group used the "Off-task/practice inquiry" entry significantly more than the high-score group (adjusted $p = .023$, d $= .662$) and low-score group (adjusted $p = .029$, $d = .552$). No significant differences were found between the high- and low-score groups.

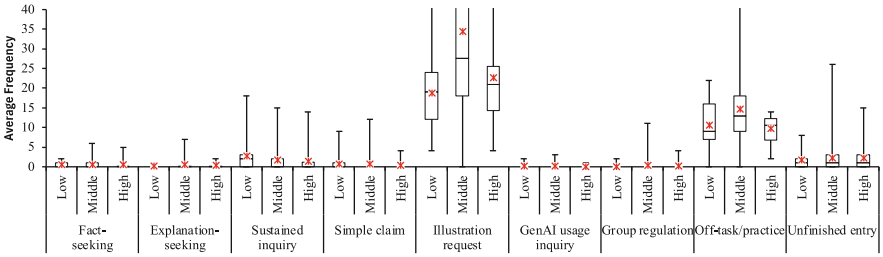


Fig. 1. Average frequency of GenAI usage code by the three groups (Boxplot)

5 Discussion

5.1 RQ1. How Do Students Use GenAI in a Collaborative Problem-Solving?

Students used GenAI much more for "Illustration request" than actions to find more innovative ideas (e.g., "fact-seeking") in this collaborative problem-solving context. One interpretation is that through using GenAI, students strengthened their recognition of GenAI that it serves more effectively as a tool for visually embodying their ideas rather than as a tool for generating innovative ideas beyond their own ideas. Meanwhile, "Prompt generation request" and "GenAI usage inquiry" were also observed, while their frequencies were low. One possible reason is that GenAI is already a user-friendly technology, without demanding users to ask prompts and how to use it.

5.2 RQ2. How Does Use of GenAI Relate to Creative Artifacts?

Our observation that frequent GenAI uses were linked with middle-level learning outcomes rather than high-level learning outcomes suggests that while GenAI can support

moderate-quality work, higher quality creative artifacts require effective integration of human creativity with AI-generated content. To identify how high-score groups achieved creative ideas, analysis of their idea development process in their discourse is needed. Notably, middle-score groups engaged more in "Off-task/practice inquiry", which may have improved their GenAI fluency and usage efficiency [16].

6 Conclusion

We conclude that while just increased use of GenAI cannot always lead to high learning outcomes, it will be a promising tool to support knowledge creation. However, students' skewed usage of GenAI indicates a need of additional support for students to use it more constructively. Further study is required to identify how high-outcome groups collaboratively developed their ideas using GenAI and what support is effective for students' more constructive use of GenAI for knowledge creation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baek, C., Tate, T., Warschauer, M.: "ChatGPT seems too good to be true": college students' use and perceptions of generative AI. Comput. Educ. Artif. Intell. **7**, 100294 (2024)
2. Cao, J., et al.: Designing an AI partner for jigsaw classrooms. In: Proceedings of the Workshop on Language-Based AI Agent Interaction with Children (AIAIC 2023) (2023)
3. Chan, C.K.Y., Hu, W.: Students' voices on generative Al: perceptions, benefits, and challenges in higher education. Int. J. Educ. Technol. High. Educ. **20**, 43 (2023)
4. Chan, C.K.K., van Aalst, J.: Knowledge building: theory, design and analysis. In: Fischer, F., Hmelo-Silver, C.E., Goldman, S.R., Reimann, P. (eds.) International Handbook of the Learning Sciences, pp. 295–307. Routledge, New York (2018)
5. Chen, B., Zhu, X., Díaz del Castillo, F.: Integrating generative AI in knowledge building. Comput. Educ. Artif. Intell. **5**, 100184 (2023)
6. Drucker, P.: The age of social transformation. Atl. Mon. **274**(5), 53–80 (1994)
7. Dwivedi, Y.K., Hughes, L., Baabdullah, A.M., et al.: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inf. Manage. **71**, 102642 (2023)
8. Jeong, H., Hmelo-Silver, C.E., Jo, K.: Ten years of computer-supported collaborative learning: a meta-analysis of CSCL in STEM education during 2005–2014. Educ. Res. Rev. **28**, 100284 (2019)
9. Lee, G.-G., Mun, S., Shin, M.-K., Zhai, X.: Collaborative learning with artificial intelligence speakers: pre-service elementary science teachers' responses to the prototype. Sci. Educ. **34**(4), 847–875 (2025)
10. Miyake, N., Kirschner, P.A.: The social and interactive dimensions of collaborative learning. In: Sawyer, R.K. (ed.) The Cambridge Handbook of the Learning Sciences, 2nd edn, pp. 418–438. Cambridge University Press, Cambridge (2014)
11. Paavola, S., Lipponen, L., Hakkarainen, K.: The knowledge creation metaphor–an emergent epistemological approach to learning. Sci. Educ. **14**(6), 535–557 (2005)

12. Ravšelj, D., Keržič, D., Tomaževič, N., et al.: Higher education students' perceptions of ChatGPT: a global study of early reactions. PLoS ONE **20**(2), e0315011 (2025)
13. Scardamalia, M., Bereiter, C.: Knowledge building and knowledge creation. In: Sawyer, R.K. (ed.) The Cambridge Handbook of the Learning Sciences, pp. 385–405. Cambridge University Press, Cambridge (2022)
14. Sublime, J., Renna, I.: Is ChatGPT massively used by students nowadays? A survey on the use of large language models such as ChatGPT in educational settings. arXiv preprint arXiv: 2412.17486 (2024)
15. Tong, Y., Chan, C.K.K.: Promoting knowledge building through meta-discourse and epistemic discourse understanding. Int. J. Comput.-Support. Collab. Learn. **18**, 353–391 (2023)
16. Toprani, D., Borge, M.: Ecological analysis of technology use among children in collaborative design contexts. J. Comput. Assist. Learn. **41**, e13119 (2025)
17. Yilmaz, R., Yilmaz, F.G.K.: Augmented intelligence in programming learning: examining student views on the use of ChatGPT for programming learning. Comput. Hum. Behav. Artif. Hum. **1**(2), 100005 (2023)



Leveraging Multi-modality and Collaborative Filtering for Supporting Automatic Scoring in Mathematics Education

Hai Li[1(✉)], Wanli Xing[1], Wangda Zhu[1], Chenglu Li[2], Bailing Lyu[2], Zifeng Liu[1], and Neil Heffernan[3]

[1] University of Florida, Gainesville, FL, USA
{li.ha,wanli.xing,wangdazhu,liuzifeng}@ufl.edu
[2] University of Utah, Salt Lake City, UT, USA
{chenglu.li,u6057059}@utah.edu
[3] Worcester Polytechnic Institute, Worcester, MA, USA
nth@wpi.edu

Abstract. This paper introduces a novel multi-modal automated scoring framework that integrates the zero-shot feature extraction of generative AI models with collaborative filtering and lightweight machine learning techniques. By encoding student responses into unified high-dimensional embeddings, our method captures both semantic and visual cues without the computational overhead of fine-tuning large-scale generative models. We then augment these embeddings with a collaborative filtering module to leverage historical similarities among student responses, thereby enhancing scoring accuracy and supporting adaptive, personalized feedback. Experimental evaluations on the ASSISTments dataset demonstrate robust predictive performance across a suite of metrics underscoring the efficacy of this multi-modal approach. Our findings highlight the transformative potential of zero-shot multi-modal representations in automated mathematics assessment and pave the way for further exploration of emerging multi-modal technologies in adaptive learning systems.

Keywords: automatic scoring · multi-modality · collaborative filtering

1 Introduction

Automated tools play a crucial role in supporting teachers by delivering consistent, efficient, and personalized feedback [12,21,24]. With access to extensive current and historical educational data, these systems can provide actionable insights to enhance student learning and assist educators in decision-making. However, the automatic assessment of open-ended responses remains challenging due to the vast diversity of student expressions, which renders traditional statistical methods [4,10] insufficient.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 313–320, 2025.
https://doi.org/10.1007/978-3-031-99264-3_39

Recent advances in artificial intelligence in education have shifted the focus towards leveraging machine learning for automated scoring [1,6,23]. Previous work on automated feedback in mathematics often relied on computationally intensive fine-tuning using specific datasets [14,22]. In particular, the zero-shot feature representation capability of generative AI (GAI) models enables the extraction of rich multi-modal features without extensive fine-tuning. This breakthrough allows us to move beyond conventional text-based natural language processing approaches [3,7,23], especially in domains such as mathematics learning where student responses often integrate textual and visual elements (e.g., graphs, diagrams, handwritten work) [11,13,27].

In this study, we introduce a novel multi-modal automated scoring framework that exploits GAI's zero-shot feature extraction in combination with traditional machine learning models. Our approach achieves robust baseline performance, which is further enhanced by integrating a collaborative filtering strategy that leverages historical multi-modal student data. Specifically, this paper addresses the following research questions:

- **RQ1**: How can multi-modal generative AI zero-shot embeddings be leveraged to assess and provide feedback on students' responses to open-ended mathematical questions?
- **RQ2**: Can the incorporation of a collaborative filtering method enhance the performance of the multi-modal automated scoring model?

2 Related Work

Our review of the literature spans two interrelated domains: GAI-based multi-modal feature representation and automated scoring methods. These areas form the foundation for our similarity-based recommendation strategy, which ultimately supports personalized learning through adaptive assessment and real-time feedback.

Conventional approaches typically extract modality-specific features, often overlooking the importance of cross-modal alignment [25]. Recent advances, however, have shifted towards embedding images and text into a unified semantic space via cross-modal deep learning [5]. Transformer-based models such as CLIP [18], BLIP [15], Mini-GPT4 [26], and ImageBind [8] have demonstrated robust capabilities in this area. In particular, CLIP and BLIP have been effectively integrated into educational feedback systems [2,12], providing a strong basis for our multi-modal feature extraction.

Automated scoring of open-ended responses has been widely explored in text-based assessments [4,5,10]. Yet, multi-modal grading remains in its nascent stage. Early methods relied heavily on manual feature engineering [9], whereas recent work leverages pre-trained cross-modal encoders to achieve unified semantic alignment [2,12]. For instance, the CLIP-Image framework has been used for similarity-based scoring, and BLIP-Double for integrating image and text in multi-label classification. Despite these advances, the application of GAI's

zero-shot feature representation in a multi-modal scoring framework is still rare. This approach offers significant advantages by reducing computational resource requirements, enabling faster inference, and seamlessly integrating with optimization techniques such as collaborative filtering—addressing a crucial gap that our work aims to fill.

3 Dataset

In this study, we utilized a dataset from online mathematics learning on the ASSISTments platform in the US. This dataset comprises open-ended mathematical questions, student responses to these questions, scores provided by teachers, and textual feedback on these answers. The mathematical questions are predominantly multi-modal, incorporating both text and visuals. Student responses come in two forms: pure images and multi-modal images with text. For students who only upload images, their text responses are automatically filled in by the online platform in the format of the image file. We obtained 12,714 responses from 3,932 students, covering 1,774 questions, with assessments provided by 513 teachers. Two attributes of responses are noteworthy: grade (mean = 5.96, std = 2.33), representing student grade levels, and difficulty level (mean = 3.78, std = 1.32), which is graded on a scale of 1 to 5 based on response time. These attributes will be utilized in the error analysis section. Response scores are distributed across five levels: 4, 3, 2, 1, 0, comprising 30.65%, 22.06%, 25.21%, 12.43%, and 9.64% respectively, indicating a lower proportion of low scores. On average, each question receives 7.16 responses, with a standard deviation of 3.73.

4 Methods

Our framework leverages the zero-shot feature representation of GAI models to extract rich multi-modal features from a diverse dataset of student responses comprising both textual explanations and visual elements (e.g., graphs, diagrams), and integrates these features with traditional machine learning and collaborative filtering techniques for robust automatic scoring.

4.1 Collaborative Filtering

To fully exploit the potential of historical data from educational platforms, we first extract embedding representations of responses to the same open-ended mathematical question and then compare their similarity. Each response is embedded in a high-dimensional vector space and compared using multiple similarity metrics—Cosine, Euclidean, and Canberra distances—all of which have proven effective in educational applications [17]. By selecting both the Top-1 most similar pair and the Top-3 most similar pairs, our method leverages not only the closest match (Top-1) but also a broader context of moderately similar cases (Top-3) to enhance prediction accuracy—an approach consistent with recommendation systems [16].

4.2 Multi-modal Feature Extraction

We extract image-text features using four transformer-based models: CLIP, BLIP, Mini-GPT4[1], and ImageBind, which produce embeddings of 512, 768, 4096, and 2048 dimensions, respectively. As previously noted, these multi-modal models include well-established baselines such as CLIP and BLIP as well as recent advancements like Mini-GPT4 and ImageBind. These multi-modal features provide rich semantic and visual cues that not only enhance the automated scoring model directly but also serve as the foundation for our similarity measures in the collaborative filtering module.

4.3 Automatic Scoring Model

New student responses are encoded into feature vectors and compared with historical data using the Canberra distance. Preliminary scores are obtained either from the best-matched response (Top-1) or through a weighted aggregation of the top three most similar responses (Top-3). These scores, along with extracted feature representations and outputs from the Item Response Theory (IRT) model, are then used as input for subsequent prediction tasks in Fig. 1. To this end, we integrate a suite of machine learning models—including Logistic Regression, Random Forest, XGBoost, and TabTransformer—within a 5-fold cross-validation framework. These models collectively cover a spectrum of learning paradigms, ranging from conventional linear classifiers and ensemble-based decision trees to deep learning architectures tailored for tabular data. In contrast to methods that rely on fine-tuning generative AI models, our approach employs lightweight classifiers, thereby reducing computational costs during both training and inference. Additionally, collaborative filtering features are explicitly incorporated to further enhance model performance. Model evaluation is conducted using Area Under the Curve (AUC), Root Mean Squared Error (RMSE), and Cohen's Kappa—metrics commonly used in automated scoring systems. Specifically, AUC reflects the model's ability to capture ordinal relationships, Kappa quantifies the agreement between predicted and true scores, and RMSE offers an interpretable measure of prediction error in terms of score deviation. To ensure rigorous evaluation, we conducted hyperparameter tuning using a sliding window approach. Key hyperparameters were adjusted, including the learning rate for all four models, the number of trees for RandomForest and XGBoost, and the depth for TabTransformer. These parameters were varied between one-quarter and four times their default values, with the optimal setting selected based on performance metrics.

4.4 Benchmarking with IRT

To ensure fairness and account for extraneous factors such as student ability, item difficulty, and rater severity, we adopt an Item Response Theory (IRT) baseline

[1] OpenAI has not provided multi-modal embedding capabilities for GPT-4o, preventing us from comparing its performance.

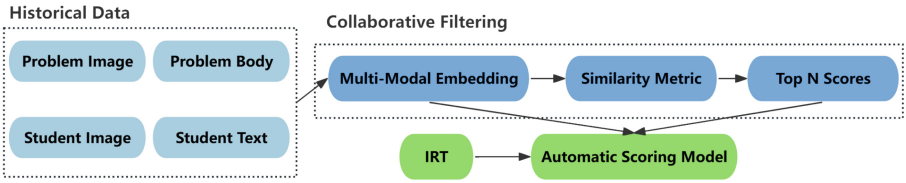


Fig. 1. The multi-modal automated scoring framework.

using the Rasch model [19,20]. This approach is widely used in the design of automated scoring models in education, and incorporating IRT properties into our evaluation enables a more equitable comparison of model performance.

5 Results

Table 1 summarizes our experimental outcomes on a dataset of multi-modal student responses to open-ended mathematical questions. We report only the results based on the Canberra distance, as it yielded the best performance compared to Cosine and Euclidean metrics in most configurations. Beginning with an IRT model paired with CLIP (ID = 1), we augment the baseline by adding image size as a predictor (ID = 2) to account for the influence of longer responses. Replacing CLIP with BLIP (IDs = 3 and 4) confirms that incorporating image size enhances performance. Further experiments with Mini-GPT4 (ID = 5) and ImageBind (ID = 6) identify BLIP and ImageBind as the superior feature extractors.

The integration of a Canberra distancebased recommendation strategy comparing Top-1 versus Top-3 selections (IDs = 710), demonstrates that incorporating moderately similar samples (Top-3) substantially enhances prediction accuracy, with ImageBind yielding optimal results. Building on these findings, we replace the baseline Logistic Regression with more complex models: RandomForest (IDs = 11, 14), XGBoost (IDs = 12, 15), and TabTransformer (IDs = 13, 16). Notably, the TabTransformer model with BLIP features and a Top-3 recommendation (ID = 13) achieves the highest AUC (80.4%) and lowest RMSE (1.27), establishing it as our best-performing model—even when compared to the higher-dimensional ImageBind variant (ID = 16). The best observed Cohen's Kappa is approximately 0.5; while values above 0.4 indicate an acceptable level of agreement, exceeding 0.6 would be considered strong performance, suggesting that further improvements are possible.

In summary, our framework's effectiveness for automated scoring in mathematics is supported by the zero-shot capability of GAI models for rich multimodal feature extraction, the enhanced feature space provided by collaborative filtering, and the resource efficiency of traditional classifiers. This approach is especially beneficial when small- to medium-sized datasets do not offer enough data to effectively fine-tune large-scale models, and it shows promise for applications in other domains involving multi-modal data and historical patterns.

Table 1. Model performance for the auto-scoring. *IRT models also included the image size as a predictor.

ID	Components of the Model	AUC	RMSE	Kappa
1	IRT + CLIP	0.6308	1.6161	0.1829
2	*IRT + CLIP	0.6312	1.6154	0.1829
3	IRT + BLIP	0.6451	1.5840	0.2289
4	*IRT + BLIP	0.6451	1.5850	0.2308
5	*IRT + MiniGPT4	0.6259	1.6458	0.1116
6	*IRT + ImageBind	0.6481	1.6181	0.2009
7	*IRT + BLIP + Canberra Top 1	0.6743	1.4942	0.3282
8	*IRT + BLIP + Canberra Top 3	0.6830	1.4649	0.3573
9	*IRT + ImageBind + Canberra Top 1	0.6819	1.4841	0.3421
10	*IRT + ImageBind + Canberra Top 3	0.6924	1.4603	0.3646
11	*IRT + BLIP + Canberra Top 3 + RandomForest	0.7196	1.4094	0.3380
12	*IRT + BLIP + Canberra Top 3 + XGBoost	0.7899	1.3052	0.4670
13	*IRT + BLIP + Canberra Top 3 + TabTransformer	**0.8042**	**1.2786**	0.4994
14	*IRT + ImageBind + Canberra Top 3 + RandomForest	0.7146	1.4543	0.2945
15	*IRT + ImageBind + Canberra Top 3 + XGBoost	0.7935	1.3072	0.4645
16	*IRT + ImageBind + Canberra Top 3 + TabTransformer	0.8040	1.2855	**0.5032**

6 Conclusion

We introduce a novel image-text multi-modal automated scoring framework that leverages the zero-shot feature extraction capabilities of GAI models, synergistically integrated with collaborative filtering and machine learning techniques. By employing lightweight classifiers instead of extensively fine-tuning large-scale models, our approach offers robust performance, scalability, and flexibility in practical applications. Looking ahead, future research may further explore the interpretability of this automated scoring framework and its integration with emerging multi-modal technologies. Additionally, interpretability techniques could be explored to design more effective collaborative learning strategies based on assignment similarity among students, as well as to develop practical instructional strategies for applying the scoring model in classroom teaching.

Acknowledgement. This work is supported by the Learning Engineering Virtual Institute under the grant number G-23-2137070 and Jaffe Foundation under the grant number AGR00026932. Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the funding agency.

References

1. Aldriye, H., Alkhalaf, A., Alkhalaf, M.: Automated grading systems for programming assignments: a literature review. Int. J. Adv. Comput. Sci. Appl. **10**(3) (2019)
2. Baral, S., Botelho, A., Santhanam, A., Gurung, A., Cheng, L., Heffernan, N.: Autoscoring student responses with images in mathematics. Int. Educ. Data Mining Soc. (2023)
3. Baral, S., Botelho, A.F., Erickson, J.A., Benachamardi, P., Heffernan, N.T.: Improving automated scoring of student open responses in mathematics. Int. Educ. Data Mining Soc. (2021)
4. Basu, S., Jacobs, C., Vanderwende, L.: Powergrading: a clustering approach to amplify human effort for short answer grading. Trans. Assoc. Comput. Linguist. **1**, 391–402 (2013)
5. Botelho, A., Baral, S., Erickson, J.A., Benachamardi, P., Heffernan, N.T.: Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. J. Comput. Assist. Learn. **39**(3), 823–840 (2023)
6. Cavalcanti, A.P., et al.: Automatic feedback in online learning environments: a systematic literature review. Comput. Educ. Artif. Intell. **2**, 100027 (2021)
7. Ghavidel, H.A., Zouaq, A., Desmarais, M.C.: Using bert and xlnet for the automatic short answer grading task. In: CSEDU (1), pp. 58–67 (2020)
8. Girdhar, R., et al.: Imagebind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180–15190 (2023)
9. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. Proc. IEEE **103**(9), 1449–1477 (2015)
10. Leeman-Munk, S., Smith, A., Mott, B., Wiebe, E., Lester, J.: Two modes are better than one: a multimodal assessment framework integrating student writing and drawing. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 205–215. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19773-9_21
11. Li, H., Guo, R., Li, C., Xing, W.: Automated quality assessment of multimodal mathematical stories generated by generative artificial intelligence. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale, pp. 110–121 (2024)
12. Li, H., Li, C., Xing, W., Baral, S., Heffernan, N.: Automated feedback for student math responses based on multi-modality and fine-tuning. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 763–770 (2024)
13. Li, H., Xing, W., Li, C., Zhu, W., Oh, H.: Are simpler math stories better? Automatic readability assessment of gai-generated multimodal mathematical stories validated by engagement. Br. J. Educ. Technol. (2024)
14. Li, H., Xing, W., Li, C., Zhu, W., Woodhead, S.: Integrating option tracing into knowledge tracing: enhancing learning analytics for mathematics multiple-choice questions. J. Learn. Anal. 1–16 (2025)
15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900. PMLR (2022)
16. Liu, N.N., Yang, Q.: Eigenrank: a ranking-oriented approach to collaborative filtering. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 83–90 (2008)

17. Oladipupo, O.O., Olugbara, O.O.: Evaluation of data analytics based clustering algorithms for knowledge mining in a student engagement data. Intell. Data Anal. **23**(5), 1055–1071 (2019)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
19. Rasch, G.: Probabilistic models for some intelligence and attainment tests. ERIC (1993)
20. Van Schuur, W.H.: Ordinal Item Response Theory: Mokken Scale Analysis. Sage (2011)
21. Wong, K.Y., Oh, K.S., Ng, Q., Cheong, J.: Linking it-based semi-automatic marking of student mathematics responses and meaningful feedback to pedagogical objectives. Teach. Math. Appl. Int. J. IMA **31**(1), 57–63 (2012)
22. Zhang, F., et al.: Math-LLMs: AI cyberinfrastructure with pre-trained transformers for math education. Int. J. Artif. Intell. Educ. 1–24 (2024)
23. Zhang, M., Baral, S., Heffernan, N., Lan, A.: Automatic short math answer grading via in-context meta-learning. arXiv preprint arXiv:2205.15219 (2022)
24. Zhang, Y., Lin, C., Chi, M.: Going deeper: automatic short-answer grading by combining student and question models. User Model. User-Adap. Inter. **30**(1), 51–80 (2020). https://doi.org/10.1007/s11257-019-09251-6
25. Zhao, R., Zhuang, Y., Zou, D., Xie, Q., Yu, P.L.: AI-assisted automated scoring of picture-cued writing tasks for language assessment. Educ. Inf. Technol. **28**(6), 7031–7063 (2023)
26. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
27. Zhu, W., Xing, W., Lyu, B., Li, C., Zhang, F., Li, H.: Bridging the gender gap: the role of AI-powered math story creation in learning outcomes. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference, pp. 918–923 (2025)

Developing Artificial Intelligence Explanations for Technology Enhanced Learning with and for Special Educational Needs Children and Teachers

Beate Grawemeyer[1(✉)] and Manolis Mavrikis[2]

[1] Coventry University, Coventry, UK
beate.grawemeyer@coventry.ac.uk
[2] UCL Knowledge Lab, University College London, London, UK
m.mavrikis@ucl.ac.uk

Abstract. This paper describes a participatory design approach involving children with special education needs (SEN) and their teachers in the design of AI explanations for a maths learning environment. Findings from these design activities indicate that the children were particularly curious about how the AI 'learned' maths. The teachers wanted not only information about the AI algorithm but also information about the data used by the AI and ethical considerations. These findings could provide practical insights to inform further design of transparent and accessible AIED systems especially for SEN children.

Keywords: Artificial Intelligence Explanations · Special Educational Needs · Participatory Design · Intelligent Learning Environments

1 Introduction

The aim of this paper is to describe how Artificial Intelligence (AI) explanations were developed with and for special educational needs (SEN) children and their teachers.

AI in technology enhanced learning is applied in different domains including for example mathematics for typical developing children (like e.g. [6]) as well as children with SEN (e.g. [5]). Such systems include data about the student, for example their knowledge level, and possible misconceptions. This data might be contained in a learner model. It is used by the system to, for example, provide tailored feedback and/or adapt the learning content. Information based on the data can be made explicit to the student and teachers. Research on how to provide this has been done in the area of e.g. open learner models [3].

Information held by the AI in a learning environment can be made transparent to those who wish to understand the system's underlying processes and decisions. For instance, students might be interested in how the AI tracks their progress or identifies specific misconceptions. Teachers may want more details

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 321–327, 2025.
https://doi.org/10.1007/978-3-031-99264-3_40

on how the system determines when a student is struggling and why. Such explanations can foster greater trust, enhance engagement, and enable more effective pedagogical strategies.

Several researchers have highlighted the importance of providing explanations of the underlying AI in technology enhanced learning e.g. [7,8,14]. Barria-Pineda et al. [1] showed that explanations of the AI decision can enhance student's engagement.

[10] provides guidelines of how such AI explanations could look like, such as providing a global explanation of the AI, examples of the AI algorithm, comparing different outcomes or conterfactual explanations of the AI. Other research showed that a combination of social and technical AI explanations are effective in providing trustworthy explanations [11] to users with different knowledge levels of AI.

The aim of this research is to find out what kind of AI explanations children as well as teachers in SEN would like to have in order to make an informed choice to engage with the AI and also to enhance the learning experience. Teachers could use the AI explanations to help provide in-person tailored support. Additionally, users and stakeholders might want to see information about the AI's ethics, to decide whether they want to engage with the system.

This paper describes the process of how children with SEN as well as teachers and a parent were involved as co-designers of the AI explanations in technology enhanced learning.

2 Participatory Design Sessions

For the development of the AI explanations several participatory design sessions were conducted. The participatory design methodology used in this research enables a voice to SEN children and teachers for identifying what kind of information they would like to have from the AI.

A local special educational needs school in Coventry was involved in the design sessions. Different levels of educational needs children are taken care of in the school. For the participatory design sessions only high functioning children who can read and write were involved.

2.1 Exploratory Learning Environment

The exploratory learning environment Frationslab was used for the development of the AI explanations [6]. The particular version of the learning environment provides a way to students to construct Fractions using different kinds of representations.

Figure 1 shows the exploratory learning environment FractionsLab. In this version tailored tasks and basic feedback are provided at the bottom of the screen. Different representations can be used in order to answer the task.

The exploratory learning environment was developed for typical developing children. It was used in this study with the children with SEN as an example of how AI can provide tailored guidance with tasks and how its tasks are sequenced.

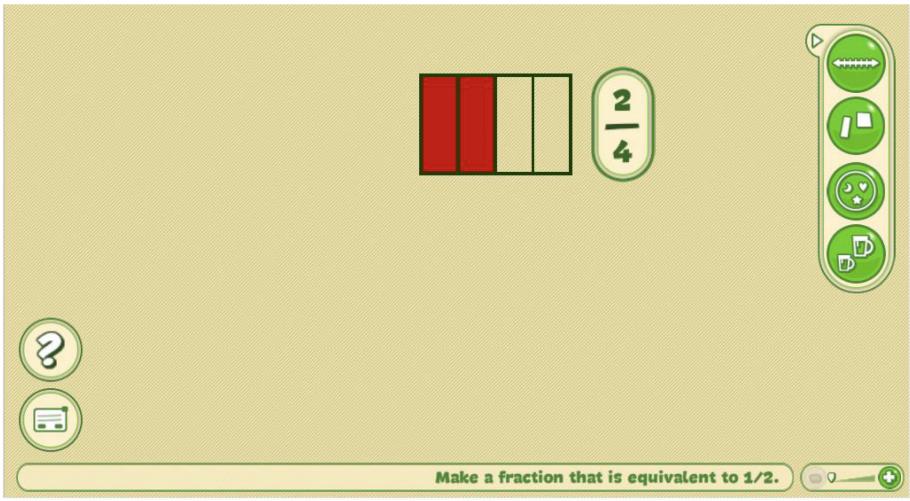


Fig. 1. The exploratory learning environment FractionsLab.

2.2 Sessions with Special Educational Needs Children

Two participatory design sessions were conducted with the children. The aim was to find out what kind of information the children would like to have from the AI in FractionsLab.

Five children with different SEN were involved in the process. All children were high functioning and could read and write and were between 10 and 12 years of age.

The children were involved in the design process similar to described in [2]. This approach is in line with broader co-design and participatory approaches positions children as active contributors to the design process [9].

At the beginning of the first session the children were able to use the learning environment for around 20 min. The purpose of this initial exploration was to serve as a grounding activity to help the children experience the system's interface, feedback mechanisms, and interactive elements.

After this they were asked to come up with explanations about the AI that they experienced during their interaction with the learning environment.

Pen and paper were used for the ideas generation. Following an open-ended prompt the children were asked to write down what they would like to ask and know about the AI in the learning environment. This session took in total 40 min.

In the second participatory design session the children were asked to refine their ideas about what kind of information they would like to have from the AI in learning environment. The session took around 20 min. The goal was to move to more concrete suggestions, identifying which explanations would be most useful for them and their peers and how they could be presented. One child was not well so there were only four children in this session.

2.3 Session with Special Educational Needs Teachers

One participatory design session with the teachers was conducted. The aim of this session was to find out what kind of information about the AI they would like to have from a learning environment that was used in their class.

Five teachers and one parent (who was also a teacher) participated in this session.

At the beginning a video was shown about the learning environment FractionsLab which explained how the AI worked. This was a slightly different version of the learning environment where also tailored feedback was provided to students based on the student's affective state. Real examples of children using the software were provided in this video too.

This was followed by ideas generation about the kind of explanations they would like to have from the AI. It was highlighted that any kind of information could be asked for including the data used, the AI algorithms or ethical issues.

As a group ideas were generated and written down on paper and discussed. This session took around 40 min.

3 Results

3.1 Special Educational Needs Children

The children came up with ideas that reflected the AI as a person that taught maths. Figure 2 shows the results of the children's second design session.

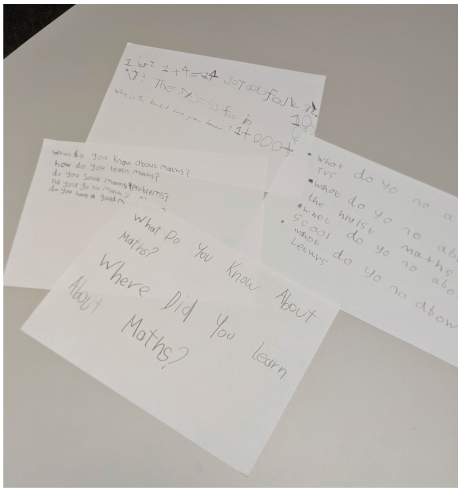


Fig. 2. The children's questions that they would like to know from the AI after using the learning environment in the second design session.

Examples of the questions that the children would like to ask the AI included for example: "How old are you?", "What do you know about maths?", "Where

did you learn about maths?", "Did you go to maths school?" or "Do you ever get questions wrong?".

It can be seen that the children were curious about the AI in the learning environment and tried to understand how the tasks were generated and the feedback provided. Their curiosity and ways they were phrasing the questions suggests that they perceived the AI as an entity and that they wanted relatable, child-friendly explanations. We return to this in the discussion.

In the second session the children refined their ideas from the first session. They suggested ways that could help them better understand the AI's decision-making process in order to increase engagement.

3.2 Special Educational Needs Teachers and Parents

The teachers and the parent were interested in different kinds of information from the AI. This included information about the data that was used and collected by the AI as well as the ethics about the AI, such as the ethical standards of the company that developed the AI etc. One participant wanted to know about how the AI worked in detail.

Examples of the information that the teachers/parent wanted to have include: "Children anonymity", "How does it (AI) work", "Ethics and information about the AI companies data is confidential and not sold", "Purpose of the AI", "training data", "How long the data is stored for".

With this kind of information the teachers and parents are able to decide if they want their student/child to engage with the learning software that includes AI.

4 Discussion and Conclusion

As described by Chaudhry et al. [4] the implementation of technology enhaced learning software should be iterative with the AI explanations at its heart.

The participatory design sessions with the children show that they were interested in getting information about the AI in the learning environment, such as how the AI obtained knowledge about maths. This indicates that the children experienced the AI in an anthropomorphic way. Research from [12] pointed out that a minimal set of human characteristics exhibited by a computer system can encourage users to exhibit human to human behaviour. The children did see the AI as a person based on the AI adaptations within the learning environment. They were also curious about how the AI did know about maths. It indicates that the children would like to know information about the AI in technology enhanced learning.

This information can be used to create AI explanations tailored to students who use learning environments with AI. Information about how maths was learned and how accurate the maths answers are can be provided.

It is interesting to see that the information that teachers would like to have is centered about data and ethics. Only one teacher wanted to know how the AI is

working in detail. This teacher was also the only one who reported high level of knowledge in computer science. This is similar to results from [11] where people with high knowledge trusted explanations about the AI more when it contained information about the AI algorithms. Whereas people with low knowledge in computer science trusted AI explanations more when they included also information about social aspects.

This indicates that not only information about the AI algorithm as described by [10] but also about data usage and the ethics behind the AI needs to be covered when providing AI explanation in technology enhanced learning in particular in special education needs settings.

It is important to include AI explanations in technology enhanced learning to teachers but as well to the students. This is particularly important for special educational needs as the AI should be as transparent as possible in order to avoid any exploitations.

The AI explanations need to be provided in a way that enables teachers and students to understand how and why the AI makes decisions including the AI algorithm, data and ethics. As described in [13] if teachers understand how the AI works then this can reduce concerns and improve trust.

Additionally, by understanding how the AI makes decisions educators can tailor their instruction more effectively and students can better engage with the learning process.

Future work includes an evaluation of different types of AI explanations for special educational needs which are based on the results of the participatory design sessions with the children and the teachers. This includes how the AI explanations should be represented (text or visual) and how long/short the AI explanations should be.

Acknowledgement. We are especially grateful to the participants who willingly gave their time to take part in this study. The authors gratefully acknowledge Sherbourne Fields Special School in Coventry. The support of REPHRAIN, UKRI (grant number EP/V011189/1) is also gratefully acknowledged.

References

1. Barria-Pineda, J., Akhuseyinoglu, K., Brusilovsky, P.: Adaptive navigational support and explainable recommendations in a personalized programming practice system. In: HT 2023: Proceedings of the 34th ACM Conference on Hypertext and Social Media (2023)
2. Benton, L., Johnson, H., Ashwin, E., Brosnan, M., Grawemeyer, B.: Developing IDEAS: supporting children with autism within a participatory design team. In: CHI 2012: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2012)
3. Bull, S., Kay, J.: Open learner models. In: Advances in Intelligent Tutoring Systems, pp. 301–322. Springer, Heidelberg (2010)
4. Chaudhry, M.A., Cukurova, M., Luckin, R.: A transparency index framework for AI in education. In: International Conference on Artificial Intelligence in Education, pp. 195–198 (2022)

5. Grawemeyer, B., Johnson, H., Brosnan, M.: Can young people with autism spectrum disorder benefit from an open learner model? In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 591–594. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19773-9_69
6. Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., Rummel, N.: Affective learning: improving engagement and enhancing learning with affect-aware feedback. User Model. User-Adap. Inter. **27**, 119–158 (2017)
7. Conati, C., Porayska-Pomsta, K., Mavrikis, M.: AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. arXiv preprint arXiv:1807.00154 (2018)
8. Conati, C., Barral, O., Putnam, V., Rieger, L.: Toward personalized XAI: a case study in intelligent tutoring systems. Artif. Intell. **298**, 10350 (2021)
9. Druin, A.: The role of children in the design of new technology. Behav. Inf. Technol. **21**(1), 1–25 (2002)
10. Khosravi, H., et al.: Explainable artificial intelligence in education. Comput. Educ. Artif. Intell. **3**, 100074 (2022)
11. Kroeger, F., Slocombe, B., Inuwa-Dutse, I., Kagimu, B., Grawemeyer, B., Bhatt, U.: Social explainability of AI: the impact of non-technical explanations on trust. In: IJCAI 2022 Workshop on Explainable Artificial Intelligence (XAI) (2022)
12. Nass, C., Steuer, J., Tauber, E., Reeder, H.: Anthropomorphism, agency, and ethopoeia: computers as social actors. In: INTERACT 1993 and CHI 1993 Conference Companion on Human Factors in Computing Systems, pp. 111–112 (1993)
13. Nazaretsky, T., Ariely, M., Cukurova, M., Alexandron, G.: Teachers' trust in AI-powered educational technology and a professional development program to improve it. Br. J. Educ. Technol. **53**, 914–931 (2022). https://doi.org/10.1111/bjet.13232
14. Woolf, B., et al.: AI-Enabled Training in Manufacturing Workforce Development. University of Massachusetts Amherst (2020)



SPAR-GNN: Knowledge Tracing with Behavioural Patterns and Selective LLM Feedback

Zhongtian Sun[1,2(✉)], Jingyun Wang[3], Ahmed Alamri[4], and Alexandra Cristea[3]

[1] University of Kent, Canterbury, UK
[2] University of Cambridge, Cambridge, UK
z.sun-256@kent.ac.uk
[3] Durham University, Durham, UK
{jingyun.wang,alexandra.i.cristea}@durham.ac.uk
[4] University of Jeddah, Jeddah, Saudi Arabia
asalamri4@uj.edu.sa

Abstract. Most knowledge tracing systems prioritise performance prediction over the delivery of personalised, actionable feedback. This paper proposes **SPAR-GNN** (Selective Pattern-Aware Reasoning with Graph Neural Networks), a novel framework that combines heterogeneous GNN-based modelling with selectively triggered feedback from a Large Language Model (LLM). SPAR-GNN represents students, problems, skills and virtual pattern nodes that encode latent behavioural traits such as frustration and hint overuse within a unified graph structure. A customised Heterogeneous Graph Layer learns student representations, while the LLM is activated only for at-risk learners identified through behavioural and performance indicators. This reduces computational overhead while enhancing pedagogical relevance. SPAR-GNN consistently outperforms strong baselines across multiple metrics, demonstrating both predictive accuracy and robustness. Observational analysis also reveals interpretable links between behaviours and learning progression, supporting fairness-aware interventions.

Keywords: Knowledge Tracing · Graph Neural Networks · Personalised Feedback · Behavioural Modelling · Large Language Models

1 Introduction

Personalised learning constitutes a foundational component of intelligent tutoring systems (ITS), with accurate *knowledge tracing* serving as a prerequisite for the delivery of tailored instructional interventions. [2,4]. However, traditional methods such as Bayesian Knowledge Tracing (BKT) and Deep Knowledge Tracing (DKT) focus primarily on correctness prediction, offering limited capacity

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 328–335, 2025.
https://doi.org/10.1007/978-3-031-99264-3_41

for interpretability or personalised feedback [3,7]. More recent approaches like Self-Attentive Knowledge Tracing (SAKT) improve model transparency but still lack mechanisms for delivering real-time, adaptive guidance [12,14].

Recent advances in Graph Neural Networks (GNNs) and Large Language Models (LLMs) present an opportunity to bridge this gap. GNNs effectively model complex relationships [9,17–20,22], but typically lack integrated feedback mechanisms. Conversely, while LLMs have exhibited considerable potential in generating educational content [11], their efficacy remains limited in predictive tasks and pedagogical interventions within educational contexts. Current evaluations suggest that LLMs demonstrate stronger capabilities in content generation than in anticipating learner needs or dynamically adapting instructional strategies to optimize pedagogical outcomes.

This paper introduces **SPAR-GNN** (*Selective Pattern-Aware Reasoning with Graph Neural Networks*), a novel framework that combines heterogeneous GNN-based knowledge tracing with behaviour-sensitive LLM feedback. SPAR-GNN incorporates virtual *pattern nodes* to represent latent behavioural archetypes, such as persistent frustration or hint overuse, enabling semantic message passing between students with similar learning challenges. Rather than fine-tuning LLMs across all interactions, SPAR-GNN selectively triggers feedback for learners identified as at risk based on predicted performance and recent behavioural indicators, enhancing both computational efficiency and pedagogical relevance.

To support fairness-aware intervention, we conduct an observational analysis of behaviourâĂŞperformance correlations, offering an interpretable and lightweight alternative to full causal inference. We evaluate SPAR-GNN on the ASSISTments 2017 dataset—a widely adopted benchmark in educational AI that features rich behavioural signals and structured skill mappings—demonstrating the framework's applicability to real-world learning contexts. Overall, SPAR-GNN advances the dual aims of knowledge tracing: delivering accurate predictions of learner progression and enabling personalised, equitable support within a scalable architecture.

2 Related Work

Bayesian knowledge tracking and deep knowledge tracking have been widely used in predicting student performance. Bayesian Knowledge Tracing models student learning as a probabilistic transition process [13], while Deep Knowledge Tracing leverages recurrent neural networks to model sequential dependencies [7]. However, these models lack interpretability and fail to generate actionable feedback [1]. Recently, GNNs have been applied in knowledge tracing to model complex relationships between students, concepts and interactions [8,21]. Although Graph-based knowledge tracking introduces student concept graphs to enhance prediction accuracy, it do not incorporate personalised feedback mechanisms [23].

LLMs, such as GPT and T5, have shown promise in generating instructional content and answering student queries [11]. Prior works have fine-tuned LLMs

for knowledge tracking [12,14], yet they incur high computational costs and lack adaptability in feedback generation. This approach selectively employs LLMs only in high-impact learning scenarios, optimising efficiency and effectiveness.

3 Methodology

3.1 Graph-Based Knowledge Tracing

This paper proposes SPAR-GNN to model student learning dynamics as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprising nodes for students, problems, and skills. To capture higher-level learning traits, we introduce virtual `pattern_nodes` representing behavioural archetypes (e.g., persistent frustration, over-reliance on hints). These nodes serve as semantic anchors, linking otherwise unconnected students through shared behavioural signals and enabling message passing across latent affective or cognitive states.

Edges encode key interactions:

- **Attempts:** (`student` → `problem`) denotes problem-solving events.
- **Skill mappings:** (`problem` → `skill`) specify assessed skills.
- **Hint-informed learning:** (`student` → `skill`) captures post-hint success.
- **Behavioural signals:** (`student` → `pattern_node`) encode latent affective or cognitive traits.

A single-layer `HGTConv` enables relation-specific message passing across this schema. Initial node features $\mathbf{h}_v^{(0)}$ are updated to $\mathbf{h}_v^{(1)}$ via attention-weighted aggregation, and final student embeddings are passed to an MLP for outcome prediction.

3.2 Selective LLM-Driven Feedback

To ensure efficiency and relevance, SPAR-GNN triggers a LLM (LLaMA-3-8B) only for at-risk students—those with predicted scores $\hat{y}_s < 0.4$ and recent behavioural signals indicating high frustration, boredom, or hint dependency. Feedback is generated using a concise prompt incorporating the student ID, dominant skill and key behavioural concerns.:

```
Student {id} is struggling with skill '{skill}' due to {issue}.
Reply with:
- A simple explanation
- One practical tip
- One real-life example
```

The resulting responses are cached and logged for transparency and reproducibility. This selective prompting strategy enables targeted, context-aware guidance while substantially reducing token usage and computational overhead, thereby improving scalability in real-world settings.

3.3 Fairness Through Behavioural Correlation Analysis

To assess fairness and potential behavioural biases, we analyse Pearson correlations between key behavioural signals (e.g., frustration levels, hint frequency) and model-predicted outcomes. For each behavioural feature x_b, we compute the standardised correlation:

$$\rho_{x_b,\hat{y}} = \frac{\mathrm{Cov}(x_b, \hat{y})}{\sigma_{x_b} \cdot \sigma_{\hat{y}}}.$$

In addition, k-means clustering is applied to the learned student embeddings to uncover latent groupings and assess whether certain behavioural profiles are disproportionately associated with low predicted performance or frequent feedback triggering. These diagnostic tools offer an interpretable lens into the fairness dynamics of the system, laying groundwork for more equitable interventions.

4 Experiments

4.1 Datasets

The ASSISTments 2017 dataset[1] [10] is derived from student interactions within the ASSISTments intelligent tutoring system. It includes a total of 942,816 problem-solving interactions from 686 unique students across 102 different questions. Among these interactions, 38.2% (359,966) were answered correctly, while the remaining 61.8% (582,850) were incorrect. The dataset records fine-grained response patterns, including timestamps, skill mappings, and behavioural indicators such as frustration and hint usage. These features make it particularly well-suited to our study of behaviour-aware knowledge tracing and selective feedback generation [5,11].

4.2 Evaluation Metrics

We evaluate all models using a comprehensive set of metrics that reflect both predictive accuracy and pedagogical utility:

- **Area Under the Curve (AUC)** âĂŞ Measures the model's ability to distinguish between correct and incorrect responses.
- **F1 Score** âĂŞ The harmonic mean of precision and recall, capturing the trade-off between false positives and false negatives.
- **Root Mean Squared Error (RMSE)** âĂŞ Quantifies the average prediction error magnitude.
- **Accuracy (Acc)** âĂŞ The proportion of correctly predicted outcomes.
- **Balanced Accuracy (Bal Acc)** âĂŞ The average of recall obtained on each class, mitigating class imbalance effects.
- **Precision** âĂŞ The proportion of predicted positives that are truly correct.
- **Recall** âĂŞ The proportion of actual positives correctly identified.

[1] https://pykt-toolkit.readthedocs.io/en/latest/datasets.html#assistments2017.

4.3 Results

Table 1 presents the performance of baseline models, including classical machine learning, LLM-based and deep learning methods. We report standard metrics: AUC, F1, RMSE, Accuracy, Balanced Accuracy, Precision and Recall. Our model, **SPAR-GNN**, consistently outperforms baselines across most metrics, highlighting its predictive strength and robustness. Results for naïve, LLM-only, and deep learning models are drawn from [11].

Table 1. Results for ASSISTments 2017 dataset. Best values in **bold**.

Family	Model	AUC	F1	RMSE	Acc	Bal Acc	Precision	Recall
Naive Baselines	Mean [15]	0.50	0.00	0.48	0.63	0.50	0.00	0.00
	NaP [15]	0.60	0.26	0.48	0.65	0.55	0.56	0.17
	NaP Skills [11]	0.59	0.35	0.51	0.63	0.56	0.50	0.27
LLM	0-Shot [11]	0.50	0.30	0.51	0.57	0.50	0.37	0.26
	FT Min [11]	0.66	0.38	0.46	0.66	0.58	0.59	0.28
	FT Extended [11]	0.68	0.39	0.46	0.67	0.59	0.63	0.28
DL	BKT [25]	0.63	0.27	0.47	0.65	0.55	0.60	0.17
	Best-LR [6]	0.70	0.48	0.45	0.69	0.63	0.62	0.39
	DKT [13]	0.77	0.58	0.42	0.72	0.68	0.64	0.54
	SAKT [12]	0.70	0.51	0.46	0.69	0.63	0.60	0.43
	SAINT+ [16]	0.54	0.01	**0.60**	0.63	0.50	0.36	0.01
Ours	SPAR-GNN	**0.845**	**0.67**	0.45	**0.80**	**0.77**	**0.65**	**0.69**

4.4 Qualitative Analysis of Selective LLM Feedback

We qualitatively assess feedback generated by the LLM for students identified as at-risk based on low predicted performance and adverse behavioural signals (e.g., high frustration or frequent hint usage). Prompts follow a consistent structure:

- **Simple Explanation**: A concise, accessible summary of the skill.
- **Practical Tip**: Behaviour-based advice tailored to the student's difficulty
- **Real-Life Example**: A relatable scenario illustrating the skill in context.

Table 2 presents representative feedback examples across diverse students and skill areas, illustrating the LLM's ability to generate personalised, context-aware guidance that addresses both cognitive and affective barriers to learning. Cognitive difficulties are supported through clear explanations and strategic scaffolding (e.g., stepwise problem-solving in Skill 27, analogy-based comprehension in Skill 26), while emotional challenges such as frustration or boredom are addressed through self-regulatory strategies like the Pomodoro technique or reduced hint

Table 2. Selected LLM-generated feedback for struggling students

ID	Skill	Feedback Summary
1825	Skill 58 (Data Analysis)	Defined data analysis, encouraged real-world connections, and illustrated with a coffee shop example
2572	Skill 27 (Problem Solving)	Advised breaking complex problems into steps, using a flat tyre to illustrate iterative problem-solving
1380	Skill 26 (Reading Comprehension)	Suggested overcoming boredom with the Pomodoro technique, and used a step-by-step cake recipe as an analogy for extracting main ideas
672	Skill 92 (Hint Dependency)	Advised reducing over-reliance on hints through incremental learning, supported by an example involving language learning and vocabulary recall
1154	Skill 43 (Data Analysis)	Encouraged contextual learning through personal interest, with an example on analysing customer data in a coffee shop for marketing strategy

reliance. By embedding suggestions in related scenarios, the feedback promotes both engagement and relevance. Overall, these results demonstrate that *selective* LLM prompting, guided by behavioural and predictive signals can provide targeted, pedagogically meaningful support without the overhead of continual inference.

4.5 Behavioural Correlates of Predicted Performance

To assess how the model aligns with student behaviour, we analyse correlations between behavioural features and predicted performance, by examining:

- **Frustration** (`RES_FRUSTRATED`): A normalised measure of the student's emotional struggle during problem-solving.
- **Hint Usage**: The average number of hints requested per student across all problem attempts.

Pearson correlations reveal a weak positive link between frustration and predicted performance ($r = 0.12$), suggesting that some frustrated students remain engaged. In contrast, hint usage shows a moderate negative correlation ($r = -0.41$), indicating that excessive reliance on hints is associated with underperformance. These findings support the model's behavioral sensitivity and motivate the use of selective interventions tailored to students' evolving emotional and cognitive states.

5 Conclusion

This paper introduced **SPAR-GNN**, a graph-based knowledge tracing framework that integrates behavioural signals and semantic pattern nodes within a heterogeneous graph. Using a customised HGT architecture and selectively triggered LLM feedback, SPAR-GNN achieves strong predictive performance (AUC = 0.845) while enhancing pedagogical efficiency. Unlike prior methods that fine-tune LLMs across all interactions, SPAR GNN invokes LLMs selectively for at-risk students, improving computational efficiency while maintaining pedagogical relevance. Observational analysis of frustration and hint usage provides a lightweight and interpretable view of fairness without requiring full causal inference. This study is limited to a single dataset (ASSISTments 2017) and one LLM backbone. Future work can explore generalisation across diverse datasets and model architectures, including those that capture behavioural and affective signals (e.g., hint usage, emotional states, engagement logs). Additionally the integration of generative approaches such as adversarial diffusion models [24], offers a promising direction for improving domain adaptablity and generating synthetic behavioural traces for low resource setting. Such extensions may also enable more principled causal reasoning and robustness across varied educational contexts.

References

1. Abdelrahman, G., Wang, Q., Nunes, B.: Knowledge tracing: a survey. ACM Comput. Surv. **55**(11), 1–37 (2023)
2. Alamri, A., Sun, Z., Cristea, A.I., Senthilnathan, G., Shi, L., Stewart, C.: Is MOOC learning different for dropouts? A visually-driven, multi-granularity explanatory ML approach. In: Kumar, V., Troussas, C. (eds.) ITS 2020. LNCS, vol. 12149, pp. 353–363. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49663-0_42
3. Alamri, A., Sun, Z., Cristea, A.I., Stewart, C., Pereira, F.D.: MOOC *next week* dropout prediction: weekly assessing time and learning patterns. In: Cristea, A.I., Troussas, C. (eds.) ITS 2021. LNCS, vol. 12677, pp. 119–130. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80421-3_15
4. Carlon, M., Cross, J.S.: Knowledge tracing for adaptive learning in a metacognitive tutor. Open Educ. Stud. **4**(1), 206–224 (2022)
5. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User Model. User-Adap. Inter. **19**, 243–266 (2009)
6. Gervet, T., Koedinger, K., Schneider, J., Mitchell, T., et al.: When is deep learning the best approach to knowledge tracing? J. Educ. Data Mining **12**(3), 31–54 (2020)
7. Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2330–2339 (2020)
8. Harit, A., Sun, Z., Yu, J., Al Moubayed, N.: Monitoring behavioral changes using spatiotemporal graphs: a case study on the studentlife dataset. In: NeurIPS 2024 Workshop on Behavioral Machine Learning (2024)

9. Harit, A., Sun, Z., Yu, J., Moubayed, N.A.: Breaking down financial news impact: a novel AI approach with geometric hypergraphs. arXiv preprint arXiv:2409.00438 (2024)
10. Liu, Z., Liu, Q., Chen, J., Huang, S., Tang, J., Luo, W.: pyKT: a python library to benchmark deep learning based knowledge tracing models. In: Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
11. Neshaei, S.P., Davis, R.L., Hazimeh, A., Lazarevski, B., Dillenbourg, P., Käser, T.: Towards modeling learner performance with large language models. arXiv preprint arXiv:2403.14661 (2024)
12. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. arXiv preprint arXiv:1907.06837 (2019)
13. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
14. Pu, S., Becker, L.: Self-attention in knowledge tracing: why it works. In: International Conference on Artificial Intelligence in Education, pp. 731–736. Springer (2022)
15. Sarsa, S., Leinonen, J., Hellas, A.: Empirical evaluation of deep learning models for knowledge tracing: of hyperparameters and metrics on performance and replicability. arXiv preprint arXiv:2112.15072 (2021)
16. Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., Choi, Y.: Saint+: integrating temporal features for ednet correctness prediction. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 490–496 (2021)
17. Sun, Z., Harit, A., Cristea, A.I., Wang, J., Lio, P.: Money: ensemble learning for stock price movement prediction via a convolutional network with adversarial hypergraph model. AI Open **4**, 165–174 (2023)
18. Sun, Z., Harit, A., Cristea, A.I., Wang, J., Lio, P.: A rewiring contrastive patch performermixer framework for graph representation learning. In: 2023 IEEE International Conference on Big Data (BigData), pp. 5930–5939. IEEE (2023)
19. Sun, Z., Harit, A., Cristea, A.I., Yu, J., Shi, L., Al Moubayed, N.: Contrastive learning with heterogeneous graph attention networks on short text classification. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2022)
20. Sun, Z., Harit, A., Yu, J., Wang, J., Liò, P.: Advanced hypergraph mining for web applications using sphere neural networks (2025)
21. Wu, Z., Huang, L., Huang, Q., Huang, C., Tang, Y.: SGKT: session graph-based knowledge tracing for student performance prediction. Expert Syst. Appl. **206**, 117681 (2022)
22. Wynn, A., Wang, J., Sun, Z., Shimada, A.: Analysing learner behaviour in an ontology-based e-learning system: a graph neural network approach (2024)
23. Yang, Y., et al.: GIKT: a graph-based interaction model for knowledge tracing. In: Hutter, F., Kersting, K., Lijffijt, J., Valera, I. (eds.) ECML PKDD 2020. LNCS (LNAI), vol. 12457, pp. 299–315. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67658-2_18
24. Yu, J., Sun, Z., Luo, S.: Adversarial diffusion model for unsupervised domain-adaptive semantic segmentation. arXiv preprint arXiv:2412.16859 (2024)
25. Yudelson, M.V., Koedinger, K.R., Gordon, G.J.: Individualized Bayesian knowledge tracing models. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 171–180. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39112-5_18



How far are we from Automatic Grading of Handwritten Cloze Form Questions?

Shrey Chandola[1(✉)], Manikandan Ravikiran[1], and Rohit Saluja[1,2]

[1] Indian Institute of Technology, Mandi, Mandi, India
s24083@students.iitmandi.ac.in
[2] BharatGen Consortium, Mandi, India
https://github.com/Shrey0900/ClozeFormHandwriting

Abstract. AI-based grading is increasingly used in education, yet its effectiveness in evaluating handwritten responses for cloze-form questions remains underexplored. This study investigates the reliability of state-of-the-art (SOTA) AI models in assessing 5,000 handwritten responses from 500 cloze-form questions, collected from students with diverse writing styles. We systematically group responses based on handwriting characteristics and compare AI-based grading with human evaluation under a strict grading policy. Our findings reveal a 29.40% average error rate for the best-performing AI model, which is 4.06% points higher than the human grader's average error rate. Additionally, AI performance varies widely across handwriting groups, indicating potential biases in automated grading systems. These results highlight the challenges of AI-based assessment in real-world educational settings and the need for robust grading models.

Keywords: Automated assessment · Cloze-form questions · OCR

1 Introduction

Assessing handwritten responses in educational settings requires significant human effort, especially in large-scale exams. AI-based grading systems have gained traction as a potential solution, particularly for standardized assessments where students provide written responses [1,12,16]. While handwriting has been linked to cognitive processes and academic performance, it remains a persistent challenge for automated assessment [3,5,15,17]. The problem is more pronounced in cloze-form questions, where answers are short and lack contextual information, making accurate evaluation difficult.

Despite advancements in AI-based grading, little attention has been given to handwriting diversity and AI biases in evaluating cloze-form responses. Existing AI-driven assessment systems often overlook variations in handwriting styles, leading to systematic grading inconsistencies. These challenges are particularly concerning in regions with limited educational resources, where AI-based assessment tools could provide scalable solutions [6,8]. However, biases in AI models could reinforce inequities rather than mitigate them.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 336–343, 2025.
https://doi.org/10.1007/978-3-031-99264-3_42

In this study, we evaluate the robustness of AI-based grading on a dataset of 500 cloze-form questions and 5,000 handwritten responses, collected from students with diverse handwriting styles. Our key contributions include:

- Comprehensive evaluation of SOTA AI models on handwritten cloze-form responses, analyzing error rate and biases across different handwriting groups.
- Comparison of AI-based and human-based grading, highlighting key discrepancies and challenges in automated assessment fairness.

Our findings reveal that even the best-performing AI model exhibits a 29.40% error rate, with notable inconsistencies across handwriting styles. This underscores the need for more adaptive, fairness-aware AI models to ensure equitable grading in educational settings.

2 Related Work

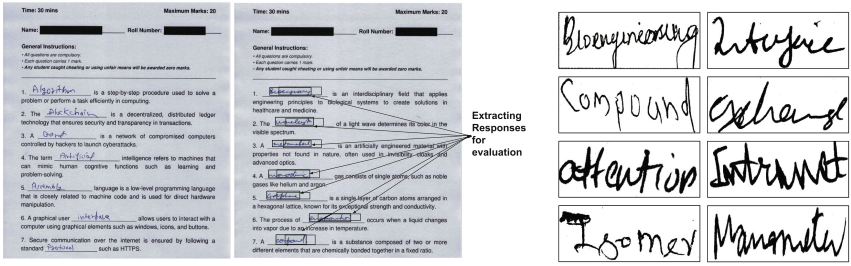
AI-Based Handwriting Recognition in Education: Handwriting recognition plays a key role in AI-driven assessment systems. Optical Character Recognition (OCR) models like TrOCR [9] and Qwen-VL [4,18] leverage deep learning and language models to better interpret handwritten text. While these models can reconstruct difficult-to-read handwriting, their accuracy often depends on writing quality, leading to grading errors [16,21]. Despite progress, limited research explores how handwriting variation affects AI performance [11,12].

AI for Cloze-Form Question Assessment: AI has been widely applied to automatically generate and evaluate cloze-form questions [10,14,20]. A system, Answers in Mind (AiM), evaluates Chinese handwritten responses but relies on examiner-provided correct answers [21]. Furthermore, AiM does not examine grading bias due to handwriting differences, which is a focus of our study.

Fairness and Bias in AI-Based Grading: Fairness in AI-based grading is critical, as models can exhibit biases based on linguistic and visual factors [7]. However, the impact of handwriting diversity on grading remains underexplored. Our work addresses this gap by analyzing how AI models struggle with specific handwriting styles, raising concerns about bias and equity in automated grading.

3 Experimental Setup

To evaluate AI-based grading in real-world educational settings, we designed an experimental framework to assess model performance on handwritten responses to cloze-form questions. Our goal is to examine how handwriting diversity affects grading accuracy and identify limitations in recognizing responses. Below, we describe the dataset, response categorization, and AI models for evaluation.



(a) (b)

Fig. 1. Overview of dataset. (a) Samples from the prototype implementation of exams. Right: Selecting responses manually for cropping. (b) Binarized version of different handwriting samples.

3.1 Prototype Exam for AI-Based Grading

We constructed a prototype exam consisting of 500 science and technology-related cloze-form questions with 5,000 handwritten responses, collected from 250 students across 9 regions of India, each providing single-word answers in designated spaces. We chose India for its linguistic diversity and wide variation in handwriting styles [13]. Each answer sheet was scanned at 600 dpi, preserving handwriting details to ensure high-quality OCR recognition. Responses were then manually cropped to remove margins and background noise before being input into AI models for text extraction. Figure 1a illustrates the structured exam format, while Fig. 1b highlights handwriting variations across responses. This structured exam format ensures controlled evaluation conditions while maintaining handwriting variability, making it suitable for AI-based grading systems.

3.2 Dataset Construction and Handwriting Categorization

To analyze AI susceptibility to handwriting diversity, we grouped all 5,000 responses into six handwriting categories based on features such as legibility, stroke variations, and letter distortions [7]. This categorization helps evaluate how readability affects AI grading performance across responses. The dataset includes students from diverse educational backgrounds, regional writing styles, and a balanced gender distribution, ensuring that our findings reflect real-world classroom scenarios and highlight fairness concerns in AI-based grading. Besides all handwritten responses were collected with informed consent from students, ensuring anonymity and compliance with ethical guidelines for human data collection in educational research.

Table 1. Performance of AI models on handwritten responses.

AI Model	Precision ↑	Recall ↑	F1-score ↑	Error Rate (%) ↓
Qwen2-VL-7B	**51.39**	**51.81**	**51.60**	**29.40**
Deepseek-vl2-small (2.8B)	42.70	45.37	44.00	36.42
LLaMA-3.2-11B-Vision	29.87	30.36	30.11	42.16
TrOCR-Large-Hw (608M)	14.47	26.07	18.61	49.30

3.3 AI Models for Handwriting Recognition

To assess how well AI models recognize handwritten responses, we evaluated four models: three vision-language models and one handwriting-specific OCR model. Vision-language models can leverage contextual cues to interpret unclear characters, making them useful for noisy handwriting. We selected Qwen2-VL-7B, Deepseek-vl2-small (2.8B), and LLaMA-3.2-11B-Vision as representative vision-language models [2,4,18,19], and TrOCR-Large-Hw (608M) as a dedicated OCR model for handwritten text [9]. Qwen2-VL-7B is pretrained on diverse OCR data sources, Deepseek-vl2 offers a lightweight alternative, and LLaMA-3.2-11B-Vision benefits from large-scale image-text pretraining. TrOCR, in contrast, is specialized for handwriting recognition, making it a strong baseline.

4 Evaluation of AI-Based Grading

We evaluate the reliability of AI-based grading for handwritten responses, analyzing recognition accuracy, bias across handwriting groups, and comparison with human evaluators. We use Precision, Recall and F1-score for character recognition and error rates for grading. Our results reveal that current AI models struggle with handwriting diversity, leading to notable grading inconsistencies.

4.1 Overall AI Model Performance

Table 1 presents the precision, recall, and F1-score to evaluate OCR recognition quality on written answers, and the error rate to measure the percentage of total responses (correct and incorrect) that the model grades incorrectly—i.e., correct answers marked wrong or incorrect answers accepted as correct. Qwen2-VL-7B [4,18] achieves the highest F1-score of 51.60 with an error rate of 29.40%, outperforming other models. However, despite being the most effective, this model still shows substantial grading errors. The consistently high error rate across all models indicates that handwriting recognition remains a major challenge in AI-driven assessments. These results highlight the limitations of current AI models in handling handwritten responses. AI grading performance is especially affected by handwriting diversity, which we analyze in the next section.

4.2 Impact of Handwriting Diversity on Grading Accuracy

To assess how handwriting style affects AI-based grading, we categorized responses into six groups based on readability and stroke complexity [7]. Table 2 shows Qwen2-VL-7B's recognition performance across these groups, with error rates ranging from 18.97% for the most readable handwriting to 49.21% for highly stylized or cursive writing. These results reveal a strong bias: students with less legible handwriting are disproportionately penalized, raising fairness concerns. The variation suggests that AI models favor structured, standardized handwriting and struggle with unconventional styles—posing a risk of unfair evaluation, especially in high-stakes testing.

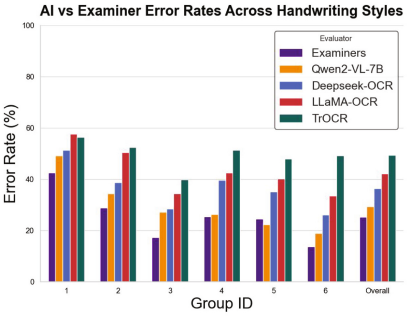
Table 2. Performance of Qwen2-VL-7B across different handwriting groups.

Handwriting Group	Precision ↑	Recall ↑	F1-score ↑	Error Rate (%) ↓
Group 1 (Least Readable)	19.60	19.80	19.70	49.21
Group 2	44.11	44.55	44.33	34.37
Group 3	55.44	55.44	55.44	27.29
Group 4	57.42	57.42	57.42	26.31
Group 5	63.46	65.34	64.39	22.23
Group 6 (Most Readable)	**68.31**	**68.31**	**68.31**	**18.97**

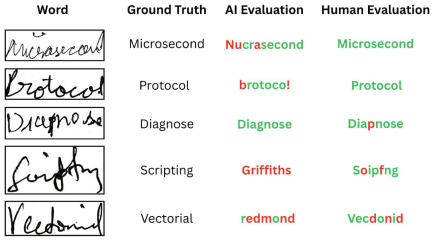
4.3 Comparison with Human Evaluation

To contextualize AI-based grading performance, we compare it with human assessments. Ten annotators from six Indian regions—including both male and female reviewers with varied educational backgrounds—conducted the human evaluation. Figure 2a shows error rates for both AI models and human evaluators. Human graders exhibit a lower error rate of 25.34%, while the best AI model (Qwen2-VL-7B) reaches 29.40%, 4.06% points higher. The gap widens for handwriting groups with high structural variability, where AI frequently misinterprets or fails to recognize words. A Wilcoxon Signed-Rank test comparing Qwen2-VL-7B and human assessments rejects the null hypothesis (no performance difference) at a 3.8% significance level, supporting our claim that humans are more reliable evaluators.

This performance gap highlights a key limitation of AI-based grading. While AI relies on character-level recognition, humans use contextual cues to adapt to stroke variations, faded writing, and overlapping characters. These findings underscore the need for human oversight in AI-driven assessments, especially for handwriting with structural variability.



(a) Error rate comparison between AI models and human examiners across handwriting groups.



(b) Representative failure cases from Qwen2-VL-7B and human grading.

Fig. 2. Evaluation of AI and human grading performance.

4.4 Error Analysis

A qualitative error analysis of representative AI-based grading outputs was conducted to understand its limitations. Figure 2b illustrates typical failure cases. In rows 1 and 2, the AI misinterprets words, highlighting its sensitivity to stroke patterns, whereas the human grader identifies the word due to familiarity with common terms. Conversely, in row 3, the AI correctly recognizes the response where the human grader misreads the handwritten input; this suggests that even human graders may occasionally struggle with visually noisy data. Rows 4 and 5 present a scenario where both AI and human graders misinterpret highly illegible handwriting, revealing the inherent difficulty of grading such responses. This analysis emphasizes the need to improve AI robustness to diverse handwriting styles and the importance of human oversight in reducing grading errors.

4.5 Challenges in AI-Based Handwriting Grading

The evaluation results reveal key challenges in AI-based grading of handwritten responses. First, AI models show strong biases toward certain handwriting styles, potentially disadvantaging students with less structured or cursive writing. Second, despite advances in OCR and vision-language models, AI still struggles with fine-grained recognition, causing frequent misclassifications. Finally, the gap between AI and human performance suggests fully automated grading is not yet feasible for handwritten cloze-form questions. These findings show that while AI-based assessment holds promise for large-scale grading, it requires major improvements in robustness, bias reduction, and contextual reasoning. We discuss future directions to address these limitations next.

5 Conclusion and Future Work

This study evaluated AI-based grading for handwritten cloze-form responses, revealing key challenges in handling handwriting diversity. Although Qwen2-VL-7B achieved the best performance, it still showed a high error rate of 29.40%, exceeding that of human graders (25.34%). AI models exhibited systematic biases, particularly struggling with stylized or cursive handwriting—raising fairness concerns in educational assessments. Future work will explore handwriting-aware AI models, context-driven post-processing, and hybrid AI-human grading approaches to improve robustness. We also plan to investigate potential biases across demographic factors such as gender, region, and handedness to ensure AI grading systems operate fairly for all students. Achieving fair, reliable, and unbiased AI grading is essential for its responsible use in education.

Acknowledgments. We thank the anonymous reviewers for their constructive feedback. This work is supported by and is part of BharatGen (https://bharatgen.tech/), an Indian Government-funded initiative focused on developing multimodal large language models for Indian languages, also funded by the Department of Science & Technology.

References

1. Deepgrade. https://smartail.ai/deepgrade/
2. AI: Llama-3.2-11b-vision. Hugging Face Model Card (2024). https://huggingface.co/meta-llama/Llama-3.2-11B-Vision
3. Asci, F., et al.: Handwriting declines with human aging: a machine learning study. Front. Aging Neurosci. **14**, 889930 (2022). https://doi.org/10.3389/fnagi.2022.889930
4. Bai, J., et al.: Qwen-vl: a versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Briggs, D.: A study of the influence of handwriting upon grades using examination scripts. Educ. Rev. **32**(2), 186–193 (1980)
6. Hammond, D., Williams, K.: Addressing the challenges of rural students in education. Edutopia (2023). https://www.edutopia.org/article/addressing-challenges-rural-students/
7. Kingrani, S.K., Levene, M., Zhang, D.: Estimating the number of clusters using diversity. Artif. Intell. Res. **7**(1), 15–22 (2018)
8. Kraus, M.W., Stephens, N.M.: Socioeconomic status and education: a complex relationship. Ann. Rev. Psychol. **64**, 1–22 (2012). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3902127/
9. Li, M., Lv, T., Cui, L., Lu, Y., Wei, F.: TrOCR: transformer-based optical character recognition with pre-trained models. arXiv abs/2109.10282 (2021). https://arxiv.org/abs/2109.10282
10. Matsumori, S., Okuoka, K., Shibata, R., Inoue, M., Fukuchi, Y., Imai, M.: Mask and cloze: automatic open cloze question generation using a masked language model. IEEE Access **11**, 9835–9850 (2023)
11. Mondal, A., Mahadevan, V., Manmatha, R., Jawahar, C.: ICDAR 2024 competition on recognition and VQA on handwritten documents. In: International Conference on Document Analysis and Recognition, pp. 426–442. Springer (2024)

12. Nguyen, H.T., Nguyen, C.T., Oka, H., Ishioka, T., Nakagawa, M.: Handwriting recognition and automatic scoring for descriptive answers in Japanese language tests. In: International Conference on Frontiers in Handwriting Recognition, pp. 274–284. Springer (2022)
13. Pal, U., Jayadevan, R., Sharma, N.: Handwriting recognition in Indian regional scripts: a survey of offline techniques. ACM Trans. Asian Lang. Inf. Process. (TALIP) **11**(1), 1–35 (2012)
14. Pino, J., Heilman, M., Eskenazi, M.: A selection strategy to improve cloze question quality. In: Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada, pp. 22–32 (2008)
15. Ruusuvirta, T., Sievänen, H., Vehkamäki, M.: Negative findings of the handwriting legibility effect: the explanatory role of spontaneous task-specific debiasing. SN Soc. Sci. **1**, 1–14 (2021)
16. Sil, P., Chaudhuri, P., Raman, B.: Can AI assistance aid in the grading of handwritten answer sheets? In: International Conference on Artificial Intelligence in Education, pp. 291–298. Springer (2024)
17. Sishwashwa, K., Mwanza, D.S.: Factors affecting the quality of handwriting among grade five learners in selected public primary schools of Mongu District, Zambia. Int. J. Res. Innov. Soc. Sci. **7**(6), 283–291 (2023). https://www.researchgate.net/publication/382463167_Factors_Affecting_the_Quality_of_Handwriting_among_Grade_Five_Learners_in_Selected_Public_Primary_Schools_of_Mongu_District_Zambia
18. Wang, P., et al.: Qwen2-vl: enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
19. Wu, Z., et al.: Deepseek-vl2: mixture-of-experts vision-language models for advanced multimodal understanding (2024). https://arxiv.org/abs/2412.10302
20. Yang, A.C., Chen, I.Y., Flanagan, B., Ogata, H.: Automatic generation of cloze items for repeated testing to improve reading comprehension. Educ. Technol. Soc. **24**(3), 147–158 (2021)
21. Zhang, Y., et al.: AIM: taking answers in mind to correct Chinese cloze tests in educational applications. arXiv preprint arXiv:2208.12505 (2022)

ProductiveMath: A Generative-AI-Powered App to Support Productive Failure Teaching

Seyedahmad Rhaimi(✉), Deniz Ercan, Ran Gao, Salah Esmaeiligoujar, Maryam Babaee, Hongming Li, Shan Zhang, Seiyon Lee, Avery Closser, and Anthony Botelho

School of Teaching and Learning, University of Florida, Gainesville, USA
{srahimi,deniz.ercan,gaoran,s.esmaeiligoujar,maryambabaee,hli3, zhangshan,leeseiyon,avery.closser}@ufl.edu, abotelho@coe.ufl.edu

Abstract. Productive Failure (PF) engages students in problem-solving before instruction but designing effective PF problems is challenging. To address this, we developed *ProductiveMath*, an AI-powered tool to support teachers in generating PF problems. Across three studies, we explored five research questions related to problem quality, AI-generation, assessment accuracy, and teacher perceptions. In Study 1, we conducted a literature review to define high-quality PF problems, created a rubric, and used GPT-4o to generate and evaluate 30 problems alongside human raters. Study 2 replicated this process with 60 additional problems, showing strong correlations between AI and human ratings, confirming GPT-4o's capability to produce high-quality problems. In Study 3, seven math teachers evaluated human- and AI-generated problems through surveys and interviews. They rated AI-generated algebra problems as high-quality, reported positive perceptions of *ProductiveMath*'s usability, and expressed intentions to use it. Key teacher feedback included suggestions to adjust problem difficulty, simplify text, enhance visuals, and provide additional PF support.

Keywords: Generative AI · Productive Failure · Algebra · K-12 · ProductiveMath

1 Introduction

In 2023, the average math score for U.S. 13-year-olds dropped 9 points from 2020, one of the largest declines since the 1970s [1]. In 2024, only 28% of eighth graders reached math proficiency, with low-performing students scoring lower than in 2022 [1]. This decline highlights gaps in foundational math skills, especially in algebra, which is crucial for academic and career success [2]. Algebra develops abstract reasoning, variable manipulation, and problem-solving, but algebra's abstract nature poses challenges [3]. Traditional instruction often disconnects algebra from daily life, limiting engagement and learning. Contextualizing algebra with real-world applications enhances its relevance and appeal [4]. Therefore, developing innovative instructional methods is essential to strengthen students' reasoning skills.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 344–351, 2025.
https://doi.org/10.1007/978-3-031-99264-3_43

The Productive Failure (PF) teaching method [4] offers a promising solution to these challenges. Failure is often perceived negatively in educational settings, yet it is a necessary component of the learning process [4]. This pedagogical strategy involves guiding students through a phase of planned failure before providing direct instructions which promotes active engagement, collaborative learning, and higher-order thinking skills [5]. Algebra is well-suited for PF due to its complexity and problem-solving opportunities [6, 7].

The PF process consists of two main phases during classroom activities: (1) Response Generation & Exploration, where students engage actively with challenging problems that typically lead to initial failures, promoting problem-solving skills and resilience; and (2) Consolidation & Knowledge Assembly, where teachers guide students toward correct solutions, reinforcing understanding. Additionally, two less-discussed phases occur before and after classroom activities: Preparation, involving teachers creating or selecting algebra problems and forming student groups, and Reflection, in which students reflect on the canonical solutions developed collaboratively with peers and teachers. Rooted in constructivist theories [8], PF posits that engaging with complex tasks before instruction helps learners restructure knowledge and build new schemas [5]. Problem-solving first connects learning to prior experiences [9], normalizes productive struggle, and fosters perseverance [10]. PF typically outperforms traditional instruction, promoting resilience, creativity, deeper learning, and knowledge transfer [7–11].

The success of PF relies on high-quality materials and effective implementation. Despite potential benefits, PF remains underused by teachers, partly because designing suitable PF problems independently is demanding and lacks extensive repositories. Effective PF problems must activate students' prior knowledge, engage them through scenarios or stories, and target a "sweet spot" of difficulty—challenging yet attainable [5]. To address this, we developed *ProductiveMath*, a generative-AI-powered application leveraging Large Language Models (LLMs) to help teachers generate PF problems. Grounded in human-centered AI principles [12, 13], *ProductiveMath* allows teachers to review, refine, and adapt AI-generated problems, supporting teacher agency and teacher-AI collaboration [14, 15]. This partnership reduces teachers' workload while maintaining instructional quality and leveraging their pedagogical expertise [16]. LLMs show strong potential in generating and evaluating educational content, particularly math word problems [17]. By adapting difficulty and context, they enable personalized instruction—crucial for PF approaches [18]. While earlier work focused on simpler problems [19], AI-generated content often requires human refinement [20]. Advanced uses include generating reading items [21], Bloom-aligned physics questions [22], creativity assessments [23], and PIRLS-style passages [24]. The three studies here address the gap in producing complex, PF-aligned math problems.

1.1 *ProductiveMath* Design Based on HCI and UX Principles

Through an iterative, rapid prototyping design process, and based on research in human-computer interaction (HCI) and user experience (UX) design (e.g. [25, 26]), the initial design of *ProductiveMath* was developed using Figma (see Fig. 2). This prototype allowed for early feedback from teachers before further development. *ProductiveMath* allows teachers to create classes and organize students into groups within the "My

Classes" page. The application was designed to gather the critical information needed for generating a PF problem in a step-by-step manner to ensure that teachers could focus on each step carefully before proceeding to the next.

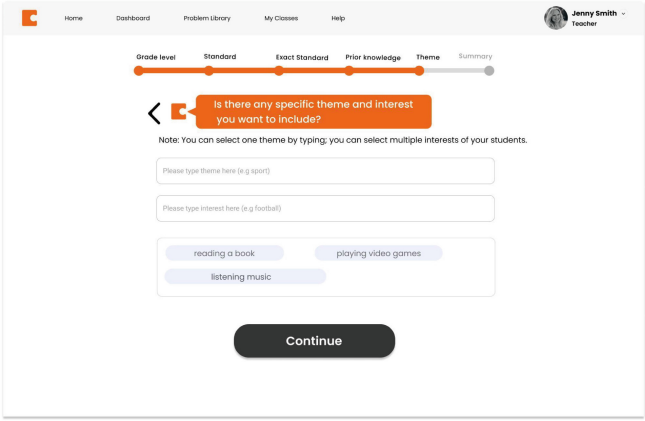


Fig. 2. *ProductiveMath*: Teacher input for AI-generated problems.

2 Study 1: Characterizing Quality for Productive Failure Content

In Study 1, we conducted a systematic literature review of algebra-related PF studies by searching databases (e.g., Google Scholar, ERIC ProQuest, Web of Science) and journals (e.g., STEM Education, Review of Educational Research). From existing PF reviews [7–11], we collected human-generated algebra problems ($n = 44$). Using seminal PF literature [10, 11], we developed a comprehensive rubric (Table 1) to evaluate problem quality, which served to train GPT-4o to generate algebra PF problems. In the second phase of Study 1, we compared zero-shot and few-shot prompt engineering approaches with GPT-4o to generate algebra PF problems. The zero-shot method provided PF guidelines without examples, whereas the few-shot method included exemplar human-generated problems from Phase 1. Iterative testing revealed that detailed contexts, such as character personas, skill descriptions, and clear task guidelines, significantly improved problem quality. The few-shot method proved most effective, generating a set of refined problems ($n = 30$). Two pairs of human raters independently assessed problem quality using the rubric, while GPT-4o simultaneously performed AI-based evaluations. Human and AI ratings were then compared to determine assessment correlation.

2.1 Results of Study 1: Establishing a Rubric for PF Problems' Quality

Our rubric (Table 1), developed from the PF literature, included five criteria: (1) Prior Knowledge Activation (connecting to students' existing knowledge), (2) Sweet Spot

Calibration (challenging yet attainable), (3) Affective Engagement (emotional engagement via scenarios), (4) Multiple Representations and Solution Methods (variety in approaches), and (5) Open-endedness and Ill-structuredness (complexity allowing multiple interpretations). Each criterion was rated on a 1–3 scale, resulting in total scores ranging from 5 to 15. The rubric was validated by a prominent PF researcher.

Human evaluations showed AI-generated problems were high-quality overall (M = 13.6, SD = 1.94), scoring well on Activation of Prior Knowledge (M = 2.88, SD = 0.37), Affective Engagement (M = 2.90, SD = 0.35), Open-endedness (M = 2.40, SD = 0.72), and Multiple Representations (M = 2.60, SD = 0.67). Pearson correlation (r = 0.78) indicated strong alignment between human and GPT-4o assessments. To further evaluate robustness, we expanded the set of PF items across multiple grade levels and standards in Study 3.

Table 1. The rubric for assessing PF problems' quality.

	1	2	3
Prior knowledge activation	Problem does not include any prerequisite concepts	Problem includes only one prerequisite concept	The problem includes 2 and more than 2 prerequisite concepts
Sweet spot calibration	Too easy or too difficult	Moderate challenge, inconsistent difficulty	Well-calibrated challenge, productive engagement
Affective engagement	Not relatable or meaningful	Some relatable elements	Highly engaging, relatable, meaningful scenario
Multiple representations and solution methods	Few or no diverse methods	Some diversity in methods	Encourages wide range of methods
Open-endedness and ill-structuredness	Simple, single solution path	Moderate complexity, few variables	Highly complex, multiple variables

3 Study 2: Exploring Automated Content Generation

In Study 2, we trained GPT-4o, through the same comprehensive prompt engineering process as we did in Study 1, to generate algebra problems suitable for PF. Specifically, in Study 2, we asked GPT-4o to generate 60 problems across 5 algebra Common Core Standards: *CCSS.Math.Content.3.OA.D.9* (n = 10), and *CCSS.Math.Content.4.OA.A.3* (n = 10), *CCSS.Math.Content.5.OA.A.1* (n = 10), *CCSS.Math.Content.7.G.B.5* (n = 15), and *CCSS.Math.Content.8.F.B.5* (n = 15). Two human raters and GPT-4o assessed the quality of those problems using the rubric.

3.1 Result of Study 2: Evaluation of LLM-Generated Content

Human raters evaluated AI-generated problems highly across rubric criteria: Activation of Prior Knowledge ($M = 2.95$, $SD = 0.19$), Sweet Spot Calibration ($M = 2.67$, $SD = 0.40$), Affective Engagement ($M = 2.96$, $SD = 0.12$), Multiple Representations ($M = 2.55$, $SD = 0.49$), and Open-endedness ($M = 2.45$, $SD = 0.52$). The overall problem quality was high ($M = 13.59$, $SD = 1.28$ out of 15), indicating GPT-4o successfully generated PF-aligned problems using iterative prompt engineering. Correlation analysis between human and GPT-4o assessments of original human-generated problems revealed strong convergent validity ($r = .78$), affirming GPT-4o's alignment with human judgment. These findings supported progression to Study 3, involving feedback from seven teachers.

4 Study 3: Teachers' Perceptions of Generated Content

The purpose of Study 3 was to compare teacher perceptions of AI-generated PF math problems to human-generated and simple problems, and evaluate the *ProductiveMath* platform. Using a mixed-methods approach [27], we collected quantitative survey ratings based on our rubric (Study 1), and qualitative data through semi-structured interviews. A participatory design method [28] guided interviews, including think-aloud sessions to gather detailed feedback on *ProductiveMath*'s usability. Seven middle-school math teachers (Grade 6 = 2, Grade 7 = 3, Grade 8 = 2) participated, recruited through district partnerships.

Teachers completed a structured Qualtrics survey evaluating three AI-generated PF problems (from *ProductiveMath* using GPT-4o), one human-generated PF problem, and one simple math problem. They rated problems independently before interviews, unaware of their origins. The survey adapted our rubric from Studies 1 and 2 into five Likert-scale items (1 = Low, 5 = High): prior knowledge activation, affective engagement, sweet spot calibration, open-endedness, and multiple representations. The survey had a maximum total score of 25. We analyzed survey data using descriptive statistics and thematic analysis [29–31].

Following the survey, teachers participated in one-hour semi-structured Zoom interviews with a live demonstration of *ProductiveMath.* Interviews explored experiences with PF, perceptions of *ProductiveMath*, usability, practical classroom applications, and improvement suggestions. Interviews were conducted by one researcher with another taking notes; notes were cross-checked for accuracy [30]. Data underwent inductive thematic analysis [31], including independent coding by researcher pairs, collaborative merging, and a final consensus meeting among five researchers, resulting in a codebook and frequency analysis.

4.1 Results of Study 3: Teacher Surveys and Interviews

Survey results indicated teachers rated AI-generated problems positively, closely matching human-generated problems (AI: $M = 17.19$, $SD = 3.07$; Human: $M = 17.43$, $SD = 3.49$). Simple problems scored lowest ($M = 13.43$, $SD = 4.20$). AI-generated problems

excelled in Sweet Spot Calibration and Affective Engagement, suggesting appropriate challenge and student engagement. However, AI-generated problems scored lower in Curriculum Alignment ($M = 2.57$, $SD = 1.43$) than human-generated problems ($M = 3.29$, $SD = 0.49$), highlighting potential improvements in aligning content with educational standards.

Qualitative data provided insights into teachers' perceptions of human- and AI-generated PF math problems. Teachers noted that human-generated problems often imposed a high cognitive load and were overly complex, especially for low-achieving students. AI-generated problems, while curriculum-aligned, sometimes lacked clarity, contained vague information, were excessively long, argumentative, or required advanced literacy skills. Despite these challenges, teachers praised AI-generated problems for effectively activating prior knowledge, fostering engagement, critical thinking, and creativity through relatable contexts (Table 2).

Table 2. Teachers' evaluation of different problem types based on the rubric. *M* (*SD*).

Problem type	*n*	PKA	SSC	AE	MR	OE	Overall	CA
Simple	7	3.29 (1.25)	2.29 (1.25)	2.43 (1.40)	3.29 (1.38)	2.00 (1.15)	13.29 (4.15)	3.29 (1.25)
Human-generated	7	3.29 (1.25)	3.14 (1.68)	3.43 (1.13)	4.00 (0.82)	3.57 (1.27)	17.43 (3.55)	3.29 (0.49)
AI-generated	21	3.00 (1.14)	3.38 (1.28)	3.71 (1.31)	3.52 (0.98)	3.57 (0.98)	17.19 (2.98)	2.81 (1.33)

Note: PKA = Prior Knowledge Activation, SSC = Sweet Spot Calibration, AE = Affective Engagement, MR = Multiple Representations and Solution Methods, OE = Open-endedness, CA = Curriculum Alignment, Overall = Average of Sum of PKA, SSC, AE, MR, and OE per problem type

Teachers recommended making AI-generated problems shorter and more concise, providing clearer instructions, adding scaffolding support, and incorporating more visual elements. Interview participants suggested improvements including visual enhancements (100%), better formatting (71%), simplified text (57%), and reading-level adjustments (42%). Specifically, teachers recommended positioning visuals clearly at the top, segmenting text with spacing and bolding key points, and aligning text complexity with students' reading abilities.

All teachers had positive initial impressions of the *ProductiveMath* prototype, describing it as "clear," "intuitive," and "user-friendly." Teachers expressed willingness to use the platform, highlighting its potential for warm-ups and small group activities, ease of problem creation and customization, and alignment with best teaching practices. Teachers also offered suggestions for enhancing *ProductiveMath*, including adding a keyword-based search feature, comprehensive support packages (solutions, misconceptions, PF guidance, answer keys), a social feature for collaboration, and integrating CLEVER for seamless account access.

5 Discussion and Conclusion

Through three studies, we evaluated GPT-4o's ability to generate Productive Failure (PF) algebra problems, informing the development of ProductiveMath—an AI-powered tool for teachers. Studies 1 and 2 showed GPT-4o produced PF-aligned problems, with some clarity and curriculum alignment issues. In Study 3, teachers valued the problems for encouraging critical thinking but suggested simplifying text, clarifying instructions, and adding visuals. They also praised ProductiveMath's design for supporting content creation and teacher-AI collaboration [23]. Future work will add student-facing features and continue educator-informed development.

References

1. National Center for Education Statistics. NAEP 2023 long-term trend assessment results. Retrieved from NCES website (2023)
2. Dougherty, S.M., Goodman, J.S., Hill, D.V., Litke, E.G., Page, L.C.: Middle school math acceleration and equitable access to eighth-grade algebra: evidence from the Wake County Public School system. Educ. Eval. Policy Anal. **37**(1), 80S–101S (2015). from https://scholar.harvard.edu
3. Common Core State Standards High School: algebra. Common Core State Standards Initiative (n.d.). https://www.thecorestandards.org/Math/Content/HSA/introduction/
4. Kapur, M.: Productive failure. Cogn. Instr. **26**(3), 379–424 (2008). https://doi.org/10.1080/07370000802212669
5. Piaget, J.: The development of thought: equilibration of cognitive structures (Trans. A. Rosin). Viking (1977)
6. Schwartz, D.L., Martin, T.: Inventing to prepare for future learning: the hidden efficiency of encouraging original student production in statistics instruction. Cogn. Instr. **22**(2), 129–184 (2004). https://doi.org/10.1207/s1532690xci2202_1
7. Sinha, T., Kapur, M.: When problem solving followed by instruction works: evidence for productive failure. Rev. Educ. Res. **91**(5), 761–798 (2021). https://doi.org/10.3102/00346543211019105
8. Darabi, A., Arrington, T.L., Sayilir, E.: Learning from failure: a meta-analysis of the empirical studies. Educ. Technol. Res. Dev. **66**, 1101–1118 (2018). https://psycnet.apa.org/doi/10.1007/s11423-018-9579-9
9. Jackson, A., Godwin, A., Bartholomew, S., Mentzer, N.: Learning from failure: a systematized review. Int. J. Technol. Des. Educ. **32**(3), 1853–1873 (2022). https://link.springer.com/article/10.1007/s10798-021-09661-x
10. Kapur, M.: Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educ. Psychol. **51**(2), 289–299 (2016). https://doi.org/10.1080/00461520.2016.1155457
11. Kapur, M., Hattie, J., Grossman, I., Sinha, T.: Fail, flip, fix, and feed–rethinking flipped learning: a review of meta-analyses and a subsequent meta-analysis. Front. Educ. **7**, 956416 (2022). https://doi.org/10.3389/feduc.2022.956416
12. Holmes, W., Bialik, M., Fadel, C.: Artificial intelligence in education: promises and implications for teaching and learning. Center for Curriculum Redesign (2019)
13. Qassrawi, R., Al Karasneh, S.M.: Redefinition of human-centric skills in language education in the AI-driven era. Stud. English Lang. Educ. **12**(1), 1894–1912 (2025)

14. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education: promises and challenges. Int. J. Educ. Technol. High. Educ. **16**(1), 1–27 (2019)
15. Kim, J.: Leading teachers' perspective on teacher-AI collaboration in education. Educ. Inf. Technol. **29**(7), 8693–8724 (2024)
16. Seo, K., Yoo, M., Dodson, S., Jin, S.H.: Augmented teachers: K–12 teachers' needs for artificial intelligence's complementary role in personalized learning. J. Res. Technol. Educ. 1–18 (2024)
17. Mohan, G.B., Prasanna, K., Vishal, K., Keerthinathan, A., Lavanya, G., Meghana, M.K.U.: An analysis of large language models: their impact and potential applications. Knowl. Inf. Syst. **66**, 5047–5070 (2024)
18. Fernández, A.A., López-Torres, M., Fernández, J.J., Vázquez-García, D.: ChatGPT as an instructor's assistant for generating and scoring exams. J. Chem. Educ. **101**(9), 3780–3788 (2024)
19. Sayin, A., Kıyak, Y.S., Kononowicz, A.A.: Using OpenAI GPT to generate reading comprehension items. Educ. Measur. Issues Pract. (2024)
20. Lin, Z., Chen, H.: Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. System **123**, 103344 (2024). https://doi.org/10.1016/j.system.2024.103344
21. Omopekunola, M.O., Kardanova, E.Y.: Automatic generation of physics items with large language models (LLMs). Res. Eval. Educ. **10**(2), 168–185 (2024). https://doi.org/10.21831/reid.v10i2.76864
22. Laverghetta Jr, A., Luchini, S., Linell, A., Reiter-Palmon, R., Beaty, R.: The creative psychometric item generator: a framework for item generation and validation using large language models. https://arxiv.org/pdf/2409.00202
23. Bezirhan, U., von Davier, M.: Generating Reading Assessment Passages Using a Large Language Model
24. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 152–155 (1994)
25. Norman, D.: The Design of Everyday Things: Revised and Expanded Edition. Basic Books (2013)
26. Design-Based Research Collective. Design-based research: an emerging paradigm for educational inquiry. Educ. Res. **32**(1), 5–8 (2003). https://doi.org/10.3102/0013189X032001005
27. Creswell, J.W., Creswell, J.D.: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage Publications (2018)
28. Spinuzzi, C.: The methodology of participatory design. Tech. Commun. **52**(2), 163–174 (2005)
29. Thomas, D.R.: A general inductive approach for analyzing qualitative evaluation data. Am. J. Eval. **27**(2), 237–246 (2006)
30. Lincoln, Y.S., Guba, E.G.: Naturalistic Inquiry. Sage Publications (1985)
31. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)



Automated Bias Assessment in AI-Generated Educational Content Using CEAT Framework

Jingyang Peng[1], Wenyuan Shen[1], Jiarui Rao[1], and Jionghao Lin[1,2(✉)]

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA
{jingyanp,wenyuan2,jiaruira,jionghal}@andrew.cmu.edu
[2] The University of Hong Kong, Pokfulam Road, Hong Kong, China
jionghao@hku.hk

Abstract. Recent advances in Generative Artificial Intelligence (GenAI) have transformed educational content creation, particularly in developing tutor training materials. However, biases embedded in AI-generated content—such as gender, racial, or national stereotypes—raise significant ethical and educational concerns. Despite the growing use of GenAI, systematic methods for detecting and evaluating such biases in educational materials remain limited. This study proposes an automated bias assessment approach that integrates the Contextualized Embedding Association Test (CEAT) with a prompt-engineered word extraction method within a Retrieval-Augmented Generation (RAG) framework. We applied this method to AI-generated texts used in tutor training lessons. Results show a high alignment between the automated and manually curated word sets, with a Pearson correlation coefficient of $r = 0.993$, indicating reliable and consistent bias assessment. Our method reduces human subjectivity and enhances fairness, scalability, and reproducibility in auditing GenAI-produced educational content.

Keywords: Fairness · Bias Detection · Large Language Models · Generative AI · CEAT · Automated Extraction

1 Background

Recent advancements in Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), have led to the development of systems aimed at assisting with educational content creation and personalized learning, demonstrating promising potential for enhancing teaching and learning [4,12,14]. However, biases embedded within AI-generated content raise substantial ethical and societal concerns, as these biases typically originate from biased or unrepresentative training datasets [5,11]. Specifically, demographic biases, which reflect societal inequalities related to characteristics such as gender, race, nationality, and ethnicity, often manifest in GenAI output by reinforcing harmful stereotypes. For example, generative AI has been shown to associate Muslims with terrorism [1]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 352–358, 2025.
https://doi.org/10.1007/978-3-031-99264-3_44

or incorrectly link Black individuals with crime [16]. Such biases compromise educational equity and further risk perpetuating existing social prejudices [6].

To systematically detect these biases, computational approaches like the Word Embedding Association Test (WEAT) have been widely employed, measuring associations between predefined sets of target words (e.g., demographic groups such as "*male*" and "*female*") and attribute words (e.g., descriptive terms such as "*career-oriented*" and "*family-oriented*") [2]. Despite its effectiveness, WEAT evaluates words in isolation without considering how meanings shift based on context, significantly limiting its ability to detect biases in realistic, context-rich educational materials. To overcome this limitation, the Contextualized Embedding Association Test (CEAT) was introduced as an extension of WEAT, incorporating contextual information by assessing word associations within actual sentences and discourses [9]. Although CEAT improves the ability to detect biases as they naturally occur in context, it still relies on manually curated word sets, introducing potential subjectivity and inefficiency in bias evaluations. To overcome the challenges associated with the manual selection of target and attribute word sets, we propose an automated method for generating context-sensitive word sets for bias assessment. Specifically, our study addresses the research question: *How does the automated extraction method compare to manually defined ground-truth word sets in CEAT-based bias evaluations?*

2 Method

2.1 Data

We used a dataset consisting of 10 AI-generated lesson scripts developed for online tutor training, many of which exhibited demographic biases (e.g., gender, racial, and national). These texts were systematically analyzed using a structured annotation process to construct ground-truth rubrics for bias identification. To evaluate the performance of our word extraction method, we also used a separate set of 4 educational texts. Full dataset descriptions, annotation protocols, and the code for automated extraction are available at Github Repository.

2.2 Bias Evaluation Using CEAT

We applied the Contextualized Embedding Association Test (CEAT) [9] to quantify bias within contextualized embeddings. Unlike static embedding tests such as WEAT [2], CEAT captures bias by considering word associations within specific textual contexts. The CEAT approach involves the following key steps:

Step 1: Computing Word Association Scores. We define two equal-sized sets of *target words* X and Y, each representing specific demographic groups (e.g. gender or nationality), and two equal-sized sets of *attribute words* A and B, which describe characteristics or qualities associated with the target groups. For each target word w, we calculate its association score as follows:

$$s(\boldsymbol{w}, A, B) = \text{mean}_{a \in A} \cos(\boldsymbol{w}, \boldsymbol{a}) - \text{mean}_{b \in B} \cos(\boldsymbol{w}, \boldsymbol{b}) \tag{1}$$

Here, $\cos(\boldsymbol{w}, \boldsymbol{a})$ represents the cosine similarity between the embedding of the word w and each attribute word a, and $\text{mean}_{a \in A} \cos(\boldsymbol{w}, \boldsymbol{a})$ stands for the average cosine similarity between the target word w and all attribute words in word set A. The difference in average cosine similarity reflects how strongly a target word w is associated with two attribute word sets, with a positive score indicating a stronger association with attributes in A and a negative score suggesting a stronger association with attributes in B.

Step 2: Calculating Effect Size (ES). The CEAT method quantifies the relative bias between two sets of target words (X and Y) using the calculated association scores. The Effect Size (ES) reflects how strongly one demographic group is associated with a particular attribute set compared to another group:

$$ES(X, Y, A, B) = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std_dev}_{w \in X \cup Y} s(w, A, B)} \tag{2}$$

The numerator in Eq. (2) captures the difference in the average association scores between two set of targets, and the denominator normalizes this difference using the standard deviation of the scores for all target words. A higher absolute value of ES indicates a stronger bias between the two demographic groups.

Step 3: Computing Combined Effect Size (CES). To summarize the bias measurements across multiple contexts, CEAT applies a random-effects model, which produces a Combined Effect Size (CES). CES is calculated as the weighted average of all individual effect sizes (ES_i) in N different contextual samples:

$$CES(X, Y, A, B) = \frac{\sum_{i=1}^{N} v_i \cdot ES_i}{\sum_{i=1}^{N} v_i}, \quad \text{where} \quad v_i = \frac{1}{\sigma^2_{\text{within}} + \sigma^2_{\text{between}}} \tag{3}$$

Here, v_i denotes the inverse of the total variance associated with each context i, considering both within-sample variance (σ^2_{within}) and between-sample variance ($\sigma^2_{\text{between}}$). The combination of various effect sizes offers a more robust measurement of bias in diverse contexts.

Step 4: Interpreting Magnitude of Bias. Finally, we interpret the resulting CES scores using Cohen's d benchmarks [15]. Values of 0.2, 0.5, and 0.8 indicate small, medium and large effect sizes respectively, helping contextualize the magnitude and potential impact of identified biases.

2.3 Automated Word Extraction

Manual creation of target and attribute sets is inefficient and subjective [9]. Therefore, we developed an automated extraction method that leverages prompt engineering within a Retrieval-Augmented Generation (RAG) framework [10].

Text Pre-processing. We split input texts into contextually coherent chunks and generated OpenAI embeddings for efficient retrieval and generation.

Few-Shot Prompting. We implemented few-shot prompting with *GPT-4o* to guide the automated extraction of demographic biases. Specifically, we designed prompts that provide clear task definitions and include illustrative examples for identifying common forms of demographic bias—namely gender, national, and racial biases. Full prompt details, including templates and constraints, are available at Github Repository.

Extraction Goals and Constraints. Prompts included explicit goals (accurate identification of demographic groups) and constraints (no inferred terms, exhaustive extraction, accurate representation of text formatting). If multiple demographic groups were identified, we computed averaged pairwise CEAT scores to ensure balanced representation.

Rubrics and Annotation Procedure. To evaluate the performance of our automated method, we created ground truth word sets through manual annotation. Three trained annotators independently reviewed each AI-generated text and extracted target and attribute word sets based on the presence of explicit or implicit demographic indicators. The annotation process followed detailed rubrics (introduced in the GitHub Repository), which ensured systematic and unbiased identification of demographic group terms and associated attributes. Any discrepancies among annotators were resolved through discussion to reach consensus. This structured annotation workflow minimized subjectivity and enabled a fair, quantitative comparison with the extracted word sets.

3 Results

3.1 Automated Word Set Extraction

To assess the performance of our automated extraction method, we systematically compared automatically generated target and attribute word sets against manually curated ground-truth sets across multiple AI-generated educational texts. As a representative example, we conducted an in-depth comparative analysis on the AI-generated course titled *"Profiles of Student Success: Inspiring Stories Across Diverse Backgrounds,"* which discusses student achievements from various cultural and demographic groups. Ground-truth word sets for this course were derived through our predefined rubric-based procedure, while the automated word sets were generated using our RAG-based method with carefully designed prompt engineering. Table 1 illustrates the comparison of these word sets, along with computed cosine similarity scores to quantitatively measure the semantic alignment between automated and ground-truth extractions.

According to Table 1, automated word sets show a strong alignment with the manually curated ground-truth sets. Target word sets from both approaches matched exactly, while attribute word sets exhibited significant overlap across all demographic groups. Cosine similarity scores ranged between 0.7627 and

Table 1. Comparison of Ground Truth and Automated Extraction Word Sets

Demographic Group	Target Words (Ground Truth)	Target Words (Extracted)	Attribute Words (Ground Truth)	Attribute Words (Extracted)	Similarity Scores
Mexican	Mexico Mexican Carlos Ramirez	Mexican Mexico Carlos Ramirez	talent perseverance discipline analytical structured environment	talent discipline perseverance analytical mind natural ability	0.7628
American	United States American Sarah Thompson	American United States Sarah Thompson	independence empathy advocacy creativity leadership social consciousness	independence creativity entrepreneurial individualism social consciousness	0.7833
Chinese	China Chinese Li Wei	Chinese China Li Wei	precision commitment academic achievement discipline methodical	discipline academic achievement precision commitment long-term goals	0.8216
Nigerian	Nigeria Nigerian Aisha Mohammed	Nigerian Nigeria Aisha Mohammed	resilience creativity empowerment innovation resourcefulness	creativity resilience empowerment breaking traditional norms innovation	0.8040
German	Germany German Michael Jensen	German Germany Michael Jensen	practicality technical expertise hands-on vocational training applied learning	hands-on practical skills technical expertise applied learning vocational training	0.8895
Indian-American	Indian-American Indian Priya Patel	Indian Indian-American Priya Patel	cultural heritage innovation holistic responsibility adaptability	tradition innovation cultural heritage modern medical practices holistic approaches	0.7627

0.8895, surpassing the 0.7 threshold indicative of strong semantic similarity [13]. Although minor differences exist, the consistently high similarity underscores the accuracy and reliability of our automated extraction method.

3.2 CEAT-Based Evaluation of Automated Extraction

To further validate our method, we compared CEAT scores computed using automated and ground-truth word sets from multiple AI-generated educational texts (Table 2). Results indicate minimal deviations between automated and ground-truth-derived CEAT scores. The highest correlation was seen in Course 3 (0.2301 vs. 0.2406), while the largest deviation occurred in Course 1 (−0.1274 vs. −0.1014), still within an acceptable margin. These results suggest that our automated extraction method effectively captures bias in AI-generated texts.

To rigorously verify the alignment between automated and ground-truth CEAT scores, we computed the Pearson correlation coefficient, which measures the strength of linear relationships between two variables [3]. Our results revealed

Table 2. CEAT Score of Ground Truth and Automated Extraction Word Sets

Text Score	Ground Truth	Automated Extraction
Tutorial Course 1	−0.1273	−0.1014
Tutorial Course 2	0.0428	0.0191
Tutorial Course 3	0.2300	0.2406
Tutorial Course 4	−0.1664	−0.1721

an exceptionally high correlation of $r = 0.9930$, indicating a near-perfect positive alignment between the two approaches. Such a strong correlation underscores that our automated extraction method reliably replicates ground-truth bias evaluations, confirming its effectiveness as a scalable and objective tool for bias assessment in AI-generated educational content.

4 Discussion and Conclusion

This study introduced an automated word extraction approach, leveraging prompt engineering within a Retrieval-Augmented Generation (RAG) framework, to systematically and objectively assess demographic biases in AI-generated educational content using GPT-4o model. The proposed method was validated through a strong Pearson correlation ($r = 0.9930$) between automated and manually curated CEAT bias evaluation scores, confirming its potential as a scalable and reliable bias assessment tool.

Educational Implications. Our approach provides an objective method for evaluating biases embedded in AI-generated educational materials, such as digital textbooks, automated learning modules, or AI-powered tutoring platforms. By proactively identifying bias, educators can enhance teaching practices by ensuring educational content is inclusive and equitable for diverse student populations. Furthermore, this method can inform policy-related initiatives aimed at regulating and monitoring the fairness and ethical standards of GenAI tools employed in educational settings.

Limitations and Future Work. Despite its promising results, our study has several limitations. First, our validation currently relies on a limited dataset of AI-generated and human-crafted educational texts; broader validation across various educational contexts and larger-scale case studies in actual classroom settings are necessary to further validate the practical effectiveness and generalizability of our approach. Second, while identifying biases is essential, effectively mitigating them is equally crucial. Our work focuses exclusively on bias detection and does not address mitigation. In the future, we should explore bias mitigation strategies across different model training phases—such as counterfactual data augmentation during pre-training [6,17], adversarial training or selective parameter freezing during model training [7,8], and human-in-the-loop approaches in the post-training phase [7].

Acknowledgment. We sincerely thank Professor Kenneth R. Koedinger for his valuable discussions and insightful suggestions on this work.

This work was fully supported by a grant from the University Research Committee (Grant No. 2401102970) at The University of Hong Kong. The opinions, findings, and conclusions expressed in this paper are solely those of the authors.

References

1. Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models (2021)
2. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
3. Cohen, I., et al.: Pearson correlation coefficient. In: Noise Reduction in Speech Processing, pp. 1–4 (2009)
4. Dai, W., et al.: Assessing the proficiency of large language models in automatic feedback generation: an evaluation study. Comput. Educ. Artif. Intell. **7**, 100299 (2024)
5. Ferrara, E.: Should chatgpt be biased? Challenges and risks of bias in large language models. First Monday (2023)
6. Ferrara, E.: Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. Sci **6**(1) (2024)
7. Gallegos, I.O., et al.: Bias and fairness in large language models: a survey (2024)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015)
9. Guo, W., Caliskan, A.: Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2021, pp. 122–133. ACM (2021)
10. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
11. Li, Y., Du, M., Song, R., Wang, X., Wang, Y.: A survey on fairness in large language models (2024)
12. Lin, J., et al.: How can i get it right? Using GPT to rephrase incorrect trainee responses. Int. J. Artif. Intell. Educ. 1–27 (2024)
13. Mikolov, T.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
14. Moore, S., et al.: Empowering education with LLMs-the next-gen interface and content generation. In: International Conference on Artificial Intelligence in Education, pp. 32–37. Springer (2023)
15. Rice, M.E., Harris, G.T.: Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law Hum. Behav. **29**, 615–620 (2005)
16. Shin, J., Song, H., Lee, H., Jeong, S., Park, J.C.: Ask LLMs directly, "what shapes your bias?": measuring social bias in large language models (2024)
17. Zmigrod, R., Mielke, S.J., Wallach, H.M., Cotterell, R.: Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. CoRR, abs/1906.04571 (2019)

Mining Student Activity Sequences in a Large-Scale Online Programming Course

Ziwei Wang(✉), Bryn Jeffries, and Irena Koprinska

School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
zwan0684@uni.sydney.edu.au, {bryn.jeffries,irena.koprinska}@sydney.edu.au

Abstract. Analysing student behaviour in online courses offers valuable insights to educators. In this paper we present methods for analysing student activity sequences to identify temporal behaviour patterns. We propose a new clustering-based approach, which integrates symbolic sequence representation, sequence similarity computation and change point detection, and also provides an interpretable visualisation of the student progress using Markov models. We conduct an evaluation in the context of a large-scale online programming course and show the effectiveness of our approach. The results enable educators to understand student behaviour patterns over time, and gain insights into the common behavioral patterns and critical points in time where student behavior changes. Our approach is applicable to other courses and provides the foundation for future research on student activity sequence analysis.

Keywords: Student activity sequence analysis · Temporal behavioural patterns · Clustering · Computer programming

1 Introduction

Online courses capture extensive data on student activities, which could help educators understand how students are engaging with the content and guide interventions to provide effective support.

Several studies have applied clustering to the sequence of actions performed by students, in order to determine temporal behavioral trends or outlier behaviors [2,3,11,13,14].

Prior work has also explored clustering student activity into a discrete set of states at different points of time and analysing the transitioning between these states via Markov modelling to understand student behavior or predict future behaviours [4,6,7,9,12]. Other studies have applied sequence mining to uncover behavioural patterns [1,13] instead of using clustering.

Our aim is to understand how student activity relates to the teaching material students engage with. For this, we need to enrich the activity sequence with information about where the student is within a course at each event. This allows us to relate behaviours to the content of a course.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 359–365, 2025.
https://doi.org/10.1007/978-3-031-99264-3_45

We propose a clustering-based approach for analysing these student activity sequences to uncover typical behavioural patterns. Our method draws inspiration from symbolic representation techniques like SAX [8] and visualisation methods such as progress networks [10], while integrating k-medoids clustering, change point detection [5] and Markov chain based visualisation.

While many prior studies focus on higher education, we apply our method to a large-scale K-12 programming course. On this scale, it is impractical to manually inspect individual student activity, and data mining methods become invaluable in helping educators identify important patterns. We aim to answer two key research questions: (i) What are the common behavioural patterns among students? (ii) Where do students commonly diverge in their behaviour?

2 Data

The dataset for this study was collected from the Beginners stream of an online Python programming course designed for high-school students and aligned with the Australian school curriculum for Years 7–8.

The course includes 10 modules. Each module consists of *content* and *problem* slides. Content slides introduce concepts and include explanations, code examples that can be run in the browser and interactive activities, e.g. asking students to modify the example code. Problem slides contain programming tasks for students to submit for automated grading and feedback. The course allows students to navigate freely between the slides without prerequisite requirements.

In this study, we used data only from the first module which consists of 15 slides - slides 0–14. Slides 4, 5, 12, 13 are problem slides and the rest are content.

A total of 11,416 students enrolled in the course. We selected all students who did at least one activity in the module, resulting in a dataset of 6,232 students. The dataset was fully anonymised, with all student identifiers replaced by randomly assigned numerical codes to ensure privacy and prevent re-identification.

3 Methodology

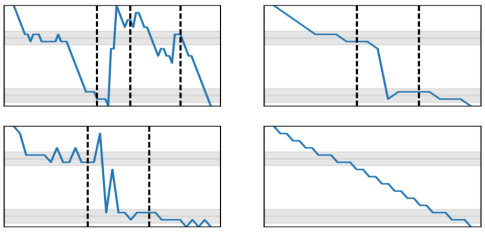
3.1 Activity Sequences and Symbolic Representation

Each student's activity is represented by a sequence of slide IDs indicating the slides they interacted with (e.g. {0, 1, 3, 6, 3, 12, 14}). The interaction includes any activity completed on the slide, e.g. slide viewed, exercises completed. In this study, we only consider the order of student interactions with the slides, and not the specific activity. As the sequence length is different for each student, we need to develop methods that can handle variable-length sequences.

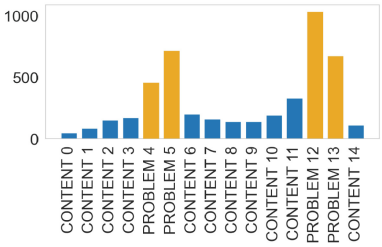
Inspired by symbolic representation methods for time series data, we developed a symbolic binning approach for the slide IDs, based on semantically meaningful labels. Traditional binning methods are typically designed for continuous data; however, as each slide contains distinct content, assuming equal distance between any two consecutive slides is inappropriate. There are 15 slides in the

module - slides 0–3 are content slides, preparing students for problem slides 4–5; similarly slides 6–11 are content slides, introducing concepts required for problem slides 12–13. Based on this semantic structure, we divided the slide IDs into five bins: Bin 0 (slides 0–3), Bin 1 (slides 4–5), Bin 2 (slides 6–11), Bin 3 (slides 12–13), and Bin 4 (slide 14).

Applying change point detection validated the appropriateness of our binning strategy. We used the Pruned Exact Linear Time (PELT) algorithm [5] to identify an arbitrary number of slope-based change points within student activity sequences. As shown in Fig. 1, the four problem slides are consistently identified as change points, supporting their use as boundaries for binning. Importantly, PELT can be applied to courses without natural semantic segmentation, providing guidance for the binning process.



(a) Examples of change points for different students. x-axis: time, y-axis: slide ID (in descending order), dashed lines: change points, white areas - content slides, grey areas: problem slides.



(b) Distribution of change points across slides, equal weight assigned to each student. Orange bars: problem slides

Fig. 1. Change point detection using PELT.

3.2 Clustering

Clustering can be used to identify meaningful student behavioural patterns and support targeted interventions. To cluster activity sequences we employed the k-medoids algorithm. It is suitable for our data due to its computational efficiency, resistance to noise, and the enhanced interpretability provided by the medoids. We also explored hierarchical clustering and density-based clustering (DBSCAN) which produced highly imbalanced clusters and spectral clustering which proved computationally inefficient for our large dataset.

To address the challenges posed by variable-length sequences, we designed a number of custom-based distance metrics based on Dynamic Time Warping (DTW), the Longest Common Subsequence (LCS), Levenshtein and Damerau–Levenshtein distances. The standard measures have limitations, e.g. the standard DTW may fail to adequately capture the true similarity as longer sequences inherently have greater distances.

We selected $D = \frac{\max(D_{LCS}) - lcs(X,Y)}{\sqrt{|X| \cdot |Y|}}$ as the distance metric for clustering, where $\max(D_{LCS})$ denotes the maximum of all LCS distances, and $lcs(X, Y)$ represents the LCS distance between sequences X and Y. This metric consistently produced well-balanced clustering results and demonstrated high computational efficiency, making it particularly suitable for large datasets. We identified this metric as the most suitable through both entropy-based evaluation and manual inspection of clustering results.

We used entropy to quantitatively evaluate the clustering results and assist in selecting the distance metric as comparing distance-based evaluation metrics like silhouette score across different measures is not meaningful. Entropy, defined as $H = -\sum_{i=1}^{k} p_i \log p_i$, where p_i is the proportion of students in the i-th cluster and k is the total number of clusters, measures the distribution of students across clusters. Higher entropy indicates balanced distributions, while low entropy signifies poor clustering with most students in a single cluster. We prioritised distance metrics that produced high mean entropy with low variance across multiple runs with different seeds, indicating stable and balanced clustering. Manual inspection further confirmed that the resulting clusters were well-defined. Additionally, our custom distance metrics demonstrated significantly better clustering performance compared to standard LCS and DTW metrics.

4 Results and Discussion

Our method produced five clusters of students, where the number of students in each cluster was: cluster 0 - 3999, cluster 1 - 441, cluster 2 - 755, cluster 3 - 666, and cluster 4 - 371. Figure 2 shows examples from each cluster.

We then constructed individual first-order Markov models for each cluster, tracking slide-to-slide transition probabilities, as shown in Fig. 3.

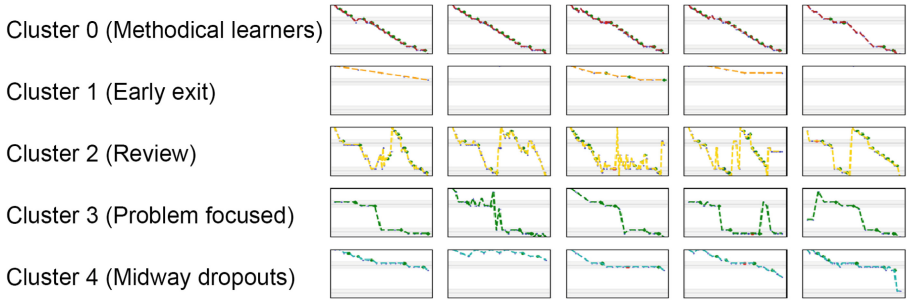


Fig. 2. Examples from each cluster. Each row represents a cluster, the first example is the cluster medoid.

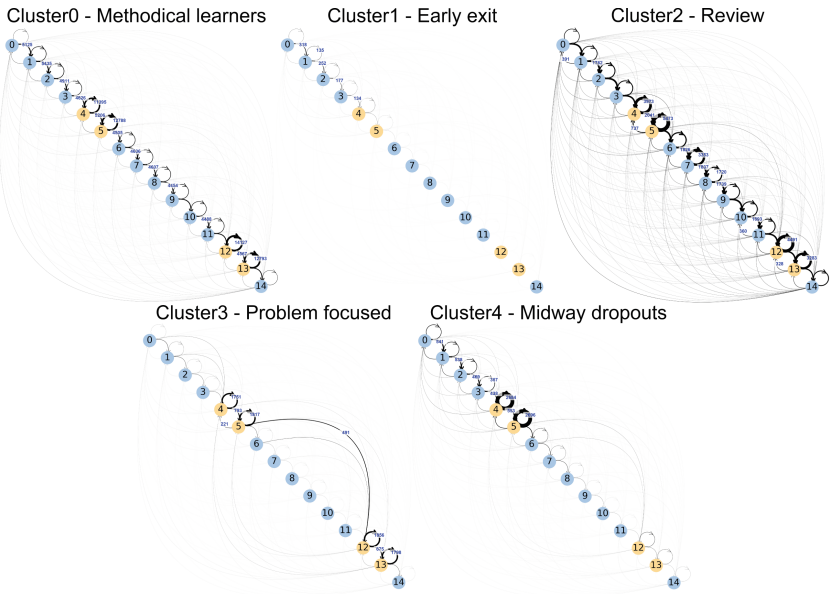


Fig. 3. First order Markov models of each cluster. Nodes represent slides, edges represent transitions with thickness reflecting frequency. Edges from the right side of a node indicate transitions to the same or later slides, while those from the left side indicate transitions to earlier slides. Orange nodes: problem slides. (Color figure online)

4.1 What are the Common Behavioural Patterns Among Students?

The clustering results reveal distinct patterns for each cluster (Fig. 3). Cluster 0, comprising 64.2% of students, is the largest group representing "Methodical learners" who consistently follow the slide sequence. Their high level of engagement and adherence to instructional design suggests that they can benefit from structured learning pathways. Cluster 1, "Early exit," represents high-risk students who disengage early, mostly before completing the first problem slide. This group requires proactive support, such as reminders or personalised assistance, to encourage continued participation. Cluster 2, "Review," exhibit a significantly higher number of interactions, with a strong tendency for extensive review after completing the module. While review can be beneficial, excessive revisits may indicate difficulties in comprehension, as seen in the repeated transitions between slides 11 and 10. Cluster 3, "Problem focused," includes students who engage mostly with problem slides, skipping content slides. This group might have prior knowledge, but could risk developing only a superficial understanding by strategically focusing only on the problem slides. Students in cluster 4, "Midway dropouts," initially engage well with the content but then disengage after facing difficulties and repeated failures. Targeted interventions, such as hints after multiple failed attempts, could help these students persist and succeed.

4.2 Where do Students Commonly Diverge in Their Behaviour?

We address this question by analysing the change point detection and clustering results. Figure 1(b) shows the distribution of change points detected, highlighting the problem slides (slides 4, 5, 12 and 13) as frequent change points. Additionally, since the clusters represent common behavioural patterns, analysing the differences between clusters reveals where students tend to diverge.

Slides 4 and 5 represent the first significant change points. With the exception of "Early exit", all other clusters have high frequency of repeated attempts on these problem slides, with "Midway dropouts" and "Review" showing the highest frequency. A backward transition from slide 5 to 4 is common, likely due to review. The "Early exit" and "Midway dropouts" exhibit minimal activity beyond these slides, marking a major disengagement. The "Problem-focused" cluster skip the intermediate content slides, jumping from slide 5 to slide 12.

Slides 12 and 13 also represent significant shifts. Similar to slides 4 and 5, they show repeated attempts and backward transitions, unlike the content slides. The "Problem-focused" and "Midway dropouts" frequently return to slides 4 and 5, possibly for review or to resume progression. Since slide 14 lacks substantial content, many "Methodical learners" and "Review" students skip it, instead returning from slides 12 and 13 directly to their intended review slides.

This demonstrates that both change point detection and clustering with Markov models for visualisation, are effective in identifying where students alter their behaviour and what these changes are. We have identified key points in the course where students disengage, review, repeat and skip content. These insights can be used by educators for tailored remedial actions to prevent disengagement and dropout and improve the learning outcomes.

5 Conclusion

This paper presents a clustering-based approach for analysing student activity sequences, to identify common temporal behaviour patterns and critical time points where student behaviour changes. Our method integrates k-medoids clustering with change point detection algorithm and Markov chain based visualisation. We designed suitable sequence representation and distance measures for sequences with variable length.

From our evaluation, using large-scale introductory programming course for high school students, we identified five clusters of distinct student behaviours: "Methodical learners", "Early exit", "Review", "Problem-focused" and "Midway dropouts". This knowledge creates opportunities for us, as educators, to cater to students with different behavioural profiles, and focus our effort more effectively. For example, problem-focused students could receive a richer set of activities, rather than spending time on explanatory content that will likely be skipped. Furthermore, the detected change-points (e.g., where students frequently repeat attempts, disengage, review, or skip content), are a valuable indicator of where we need to provide additional support to students in classroom discussions, or where content needs rewriting to make it more accessible.

In future work, we plan to extend our method to other modules, explore how students transition between clusters over longer time spans and investigate how the learning behaviours impact course performance.

References

1. Abuabat, A., Brusilovsky, P.: Sequence analysis of learning behavior in different consecutive activities. In: Proceedings of E-Learn: World Conference on EdTech, pp. 817–822 (2018)
2. Cobo, G., García-Solórzano, D., Santamaría, E., Morán, J.A., Melenchón, J., Monzo, C.: Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering. In: Proceedings of the International Conference on Educational Data Mining (2011)
3. Desmarais, M., Lemieux, F.: Clustering and visualizing study state sequences. In: Proceedings of the International Conference on Educational Data Mining (2013)
4. Hansen, C., Hansen, C., Hjuler, N., Alstrup, S., Lioma, C.: Sequence modelling for analysing student interaction with educational systems. In: Proceedings of the International Conference on Educational Data Mining (2017)
5. Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. **107**(500), 1590–1598 (2012)
6. Klingler, S., Käser, T., Solenthaler, B., Gross, M.: Temporally coherent clustering of student data. In: Proceedings of the International Conference on Educational Data Mining (2016)
7. Köck, M., Paramythis, A.: Activity sequence modelling and dynamic clustering for personalized e-learning. User Model. User-Adap. Inter. **21**, 51–97 (2011)
8. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Disc. **15**(2), 107–144 (2007)
9. Maqsood, R., Ceravolo, P., Romero, C., Ventura, S.: Modeling and predicting students' engagement behaviors using mixture Markov models. Knowl. Inf. Syst. **64**(5), 1349–1384 (2022)
10. McBroom, J., Paassen, B., Jeffries, B., Koprinska, I., Yacef, K.: Progress networks as a tool for analysing student programming difficulties. In: Proceedings of the 23rd Australasian Computing Education Conference, pp. 158–167 (2021)
11. McBroom, J., Yacef, K., Koprinska, I.: DETECT: a hierarchical clustering algorithm for behavioural trends in temporal educational data. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 374–385 (2020)
12. Polito, S., Koprinska, I., Jeffries, B.: Exploring student engagement in an online programming course using machine learning methods. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 546–550 (2022)
13. Ulitzsch, E., He, Q., Pohl, S.: Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. J. Educ. Behav. Stat. **47**(1), 3–35 (2022)
14. Zhu, M., Wang, X., Xu, Q., Andrews-Todd, J.: Behavior sequence patterns in collaborative problem solving: a multiple-perspective analysis. In: Proceedings of the International Conference on Computer-Supported Collaborative Learning, pp. 205–208 (2024)

Assessing Lecture Difficulty and Unexpected Events Using Subtle Facial Cues

Tadachika Ozono(✉), Yuna Kaneko, and Masato Kikuchi

Nagoya Institute of Technology, Nagoya, Japan
{ozono,kikuchi.masato}@nitech.ac.jp, kaneyu@ozlab.org

Abstract. This study bridges subtle facial cues and macro-engagement to improve lecture difficulty and unexpected event prediction using two-class lecture attitude metrics—r_{class} and ρ_{class}. Previous work has integrated individual facial emotion recognition (FER) outputs to estimate classroom engagement, yet the relationship between subtle cues and macro-level engagement remains unclear. We extend prior methods by incorporating soft output values to capture subtle facial cues more effectively. Evaluation experiments in simulated lectures highlight that ρ_{class}, derived from soft output probabilities, outperforms the previous metric, r_{class}, in detecting unexpected events (0.317 vs. 0.000). These results demonstrate the advantage of soft outputs in capturing subtle facial cues and their potential for real-time classroom assessment and intervention.

Keywords: Facial Expression Analysis · Classroom Engagement · Lecture Difficulty Assessment · Unexpected Event Detection

1 Introduction

Facial expression recognition (FER) offers a non-intrusive and real-time approach to detecting affective states in learning environments. Such a capability is particularly valuable in online settings, where instructors cannot observe students' expressions directly. Although moderate confusion may support productive struggle, sustained negative affect often signals disengagement. Detecting such subtle cues enables timely instructional support.

Understanding when students perceive a lecture as difficult or encounter unexpected events is crucial for adapting instruction and maintaining engagement. These situations often manifest as subtle facial expressions that are difficult to detect using categorical emotion labels alone. We build on prior work and propose a complementary metric, ρ_{class}, derived from soft FER outputs, which captures subtle emotional signals to estimate perceived lecture difficulty and identify unexpected events. Though r_{class} aggregates categorical emotion labels [9], ρ_{class} leverages soft probabilities to represent finer dynamics. Our experiments show ρ_{class} consistently outperforms r_{class} in detecting subtle shifts

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 366–373, 2025.
https://doi.org/10.1007/978-3-031-99264-3_46

in engagement, suggesting these metrics offer complementary insights into student affect. Prior work suggests that affect-aware instruction supports metacognitive regulation and attention, critical for effective learning [2].

Recent multimodal learning analytics research [1,4,7] highlights the potential of combining facial expressions with gaze and physiological data to infer cognitive and emotional states. For example, Mangaroska et al. [7] demonstrated that facial expressions can predict cognitive load during high-demand programming tasks. However, linking these cues to macro-level engagement remains challenging, especially in low-demand instructional settings like ours.

2 Related Work

Several methods have been proposed to estimate lecture attitudes based on students' behavioral responses, including facial expressions, gaze, and body movements. FER is widely used because it is low-cost and requires only a camera.

Pabba et al. [8] developed a system that classifies student engagement into three levels based on FER and achieved 73.68% accuracy in predicting self-reported motivation. Xiaoyu et al. [10] used FER and the PAD model to calculate emotional engagement, a significant predictor of performance. Swadha et al. [3] proposed a deep learning-based method for real-time engagement detection in online classes with about 90% accuracy. Mahendar et al. [6] proposed a model to recognize student micro-facial expressions as subtle emotional cues during online classes to assess emotional states and learning rates. However, factors such as lecture difficulty and unexpected events, especially those reflected through subtle facial expressions, have not been thoroughly investigated.

Buono et al. [1] explored multimodal engagement estimation using facial expressions, head orientation, and gaze, though their method showed only a weak correlation (0.39) with questionnaire-based engagement. Kok et al. [4] showed that visualized gaze data allows participants to infer test scores above chance levels. However, gaze data collection is costly and less practical for classrooms. FER-based class lecture attitudes offer a scalable, accessible alternative, though their potential for engagement inference requires further exploration.

Therefore, this study investigates how lecture difficulty and unexpected events affect class lecture attitude, an area not thoroughly explored in prior work. We approach this from two perspectives: categorical (hard) FER and soft-probability-based FER. The method involves presenting waveform data of students' class lecture attitudes to instructor subjects and asking them to infer the corresponding questionnaire responses.

3 Classroom Evaluation Methods

Class lecture attitude represents the overall emotional state of a class based on students' facial expressions during a lecture. Peng et al. [9] proposed using r_{class}, a measure based on hard facial recognition, to estimate class lecture attitude. This method determines each student's response using single-class classification

and aggregates these responses to compute r_{class}. A larger value of r_{class} corresponds to a more positive class lecture attitude, whereas a smaller value reflects a more negative attitude.

In this study, seven emotions are categorized into three polarities: positive, neutral, and negative. Specifically, $Em^+ = Happy$ (positive), $Em^0 = Neutral, Surprise$ (neutral), and $Em^- = Sad, Fear, Disgust, Anger$ (negative). We define $Em^{\text{all}} = Em^+ \cup Em^0 \cup Em^-$. A FER model returns a soft output vector $\mathbf{p}_{\text{soft}} = (p_j)$, where each p_j is the probability of emotion $j \in Em^{\text{all}}$ for a given facial image. For example, $\mathbf{p}_{\text{soft}} = (p_{Happy}, p_{Neutral}, \ldots) = (0.7, 0.3, \ldots)$.

In hard FER, r_{class} is defined as follows:

$$r_{\text{class}} = \sum_{i=1}^{n} w_i \cdot r_i, \quad r_i = 1 \text{ (for } Em^+), \ 0 \text{ (for } Em^0), \text{ and } -1.2 \text{ (for } Em^-), \quad (1)$$

where r_i represents the response of the i-th student among n students. The weights reflect the frequency of detection of non-neutral emotions for student i, giving more influence to those with consistent emotional signals; r_i values follow Peng et al. [9], penalizing negative emotions more to capture disengagement. In hard FER, the result corresponds to the emotion with the highest probability in $\mathbf{p}_{\text{soft}}$. Analysis of FER results indicates that non-neutral soft output values are slightly lower than neutral ones, complicating the detection of subtle responses via hard FER. This limitation in Eq. (1) reduces its effectiveness in capturing these nuances.

To address this issue, we introduce Eq. (2) based on soft output values:

$$\rho_{\text{class}} = \frac{1}{n} \sum_{i=1}^{n} \rho_i, \ \rho_i = \frac{\sum_{j \in Em^+ \cup Em^-} w_j \cdot p_{ij}}{\sum_{j \in Em^+ \cup Em^-} |w_j| \cdot p_{ij} + \sum_{j \in Em^0} p_{ij}}, \quad (2)$$

where ρ_i is the emotional polarity of the i-th student, and n is the total number of students. Namely, ρ_{class} is the average of ρ_i. Here, r_i can be interpreted as a student's immediate emotional polarity, while ρ_i reflects the extent to which non-neutral emotions dominate, enabling more sensitive detection of subtle class-wide shifts. In this equation, p_{ij} denotes the soft output value of emotion $j \in Em^{\text{all}}$ for the i-th student, and w_j is the weight assigned to each emotion: $+4$ for $j \in Em^+$ and -1 for $j \in Em^-$. The key advantage of ρ_{class} is its ability to capture subtle facial cues by leveraging the distribution of soft output values, enabling a refined assessment of macro-level engagement and class lecture attitudes. Importantly, it emphasizes the proportion of non-neutral emotions, providing a clearer signal when emotional responses diverge from neutrality—an aspect often missed by r_{class}. This method remains effective even when FER accuracy is limited.

4 Experiments

We examined whether common patterns exist between the time-series changes in r_{class} and ρ_{class}. The experiment was conducted in two stages: (1) dataset construction based on simulated lectures and (2) difficulty estimation and unexpected event detection by humans based on the changes of r_{class} and ρ_{class}.

4.1 Dataset Construction and Methods

We constructed a dataset of student facial expressions and questionnaire responses from simulated lectures. The dataset includes facial images, FER results $\mathbf{p}_{\text{soft}}$, and class lecture attitude based on r_{class} and ρ_{class}.

In the simulated lectures, the following steps were conducted: (a) Problem: problem introduction and thinking time, constrained to either 25 or 60 s depending on the problem, (b) Hint: hint presentation (10 s for 25-second problems, 15 sec for 60-second problems), (c) Answer: answer disclosure (5 sec, passive viewing of the correct answer), and (d) Questionnaire: a questionnaire for the subjects (30 sec). Problems varied in difficulty to elicit diverse responses. Some hints were misleading and some answers incorrect, forming predefined unexpected events as ground truth for evaluation. In this study, these instances are referred to as unexpected events. In (d) Questionnaire, the subjects were asked to respond to two questions: the perceived difficulty during (a) Problem and (c) Answer. Difficulty was rated on a five-point verbal scale: very easy, easy, moderate, hard, and very hard, scored as 1, 2, 3, 4, and 5, respectively. The subjects comprised 8 Asian students (5 undergraduates, 3 graduates; 6 males, 2 females; aged 21–24) solved 15 problems. Facial expressions were recorded using built-in laptop cameras at 30 fps and sampled at 15 fps. OpenFace[1] cropped facial regions, which were then analyzed using POSTER [11] with a model pre-trained on RAF-DB [5]. Subsequently, the temporal changes in class lecture attitude were calculated frame by frame using r_{class} and ρ_{class}.

We investigated whether the subjects acting as lecturers could estimate students' perceived difficulty from the time-series changes of r_{class} and ρ_{class}. Note that the lecturer subjects neither participated in nor observed the simulated lectures. We divided the time-series data of class lecture attitude for 15 problems into two parts: 5 for a preliminary experiment and 10 for the main experiment. First, three subjects analyzed the five preliminary problems, revealing two rules from the time-series changes of r_{class} and ρ_{class}. Later, we added three more subjects, bringing the total to six. They were given the time-series data for the remaining 10 problems (P1–P10), questionnaire results, data from the five preliminary problems as hint material, and the discovered rules. Notably, P5 had an inappropriate hint, and P10 contained an incorrect answer. The lecturer subjects estimated the difficulty of each problem and identified unexpected events. Difficulty was also rated on a five-point verbal scale: very easy, easy, moderate, hard, and very hard, corresponding to scores of 1 to 5. In addition, they reported unexpected events for each problem, specifically identifying inappropriate hints and incorrect answers.

Figure 1 presents examples of the hint materials, showing time-series changes in class lecture attitude for ρ_{class} (left) and r_{class} (right) with time on the horizontal axis and class lecture attitude on the vertical axis, along with the corresponding difficulty levels, easy (left) and hard (right). These examples illustrate two rules: Rule 1 indicates that the class lecture attitude stays around zero longer

[1] https://github.com/TadasBaltrusaitis/OpenFace.

when the problem is easy, and Rule 2 indicates that the class lecture attitude exceeds zero more frequently when the problem is easy.

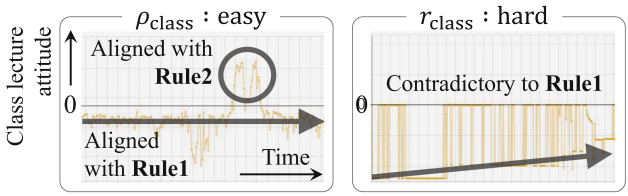


Fig. 1. Hint material with rules

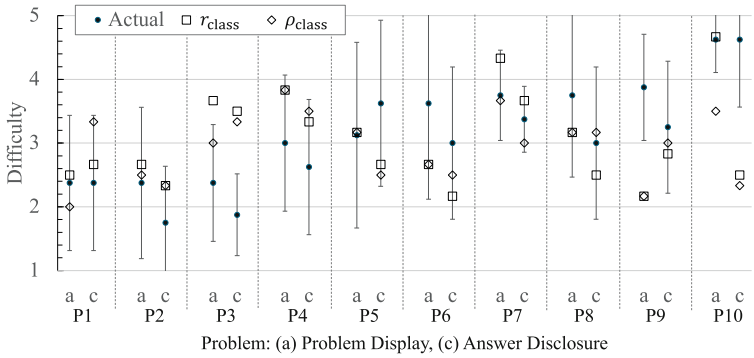


Fig. 2. Comparison of actual and predicted difficulty based on r_{class} and ρ_{class}

4.2 Results and Analysis

Figure 2 shows the comparison between self-reported difficulty (Actual), and the predicted difficulty values r_{class} and ρ_{class} as interpreted by the lecturer subjects for each problem (P1–P10). The analysis compares the performance of r_{class} and ρ_{class} at two stages: (a) problem display (P1a–P10a) and (c) answer disclosure (P1c–P10c). The dots represent the mean self-reported difficulty, with error bars indicating the standard deviation, reflecting the variability in students' subjective rating. The squares and diamonds show r_{class} and ρ_{class}, respectively.

The results reveal distinct patterns in the performance of r_{class} and ρ_{class} across different problems and stages. The lecturer subjects more accurately inferred class lecture attitudes from ρ_{class} during (c) answer disclosure, while r_{class} supported better inference during (a) problem display in some instances

(e.g., P1 and P10). These findings indicate that when interpreted by lecturer subjects, r_{class} and ρ_{class} provide complementary insights into students' self-reported difficulty, emphasizing the need to use both metrics for comprehensive analysis. For problems P5 and P10 involving unexpected events, the difficulty inferred by lecturer subjects from r_{class} and ρ_{class} exhibited notable errors in (c) answer disclosure, suggesting confusion in assessing the correct difficulty. Similarly, the predictions for P3c and P9a were also poor, though the underlying reasons for these outcomes remain unclear.

Table 1 shows the macro-average precision of the lecturer subjects for detecting unexpected events using r_{class} and ρ_{class}. The precision was 0.000 (r_{class}) and 0.317 (ρ_{class}), indicating that ρ_{class} is more effective in identifying unexpected events. Notably, ρ_{class} exceeds the chance rate of 0.200. This result suggests that soft output values enhance sensitivity to subtle changes in class lecture attitude.

One limitation of this study is the focus on problem solving rather than traditional passive listening, likely eliciting more responses. Another is the small sample size and limited ethnic diversity among participants, which may reduce the generalizability of the results. Future work includes expanding participant numbers, adding multimodal data, and real-time prototype integration. Our preliminary findings confirm soft FER outputs' potential to enhance classroom engagement assessment. While confusion or frustration can sometimes support learning through productive struggle, the proposed method aims to detect excessive or unproductive difficulty, enabling timely instructional support. For instance, if a sustained drop in ρ_{class} is detected during a lecture, instructors could slow down, offer additional explanations, or switch instructional strategies.

Table 1. Prediction Precision for Unexpected Events Detection

Evaluation Target	r_{class} Precision	ρ_{class} Precision
Unexpected Event Detection	0.000	**0.317**
Chance Rate	0.200	

5 Conclusions

This study examined how lecturer subjects predicted lecture difficulty and unexpected events using two metrics—r_{class} and ρ_{class}. We aimed to bridge the gap between subtle facial cues and macro-level engagement by incorporating soft output values into class lecture attitude analysis. The results suggest that ρ_{class} supported more accurate inferences than r_{class} in detecting subtle emotional responses, with a precision of 0.317 (vs. 0.000) for unexpected events. Given the chance rate of 0.20, these results suggest the effectiveness of soft output-based metrics in capturing complex classroom dynamics.

In addition, r_{class} and our proposed metric ρ_{class} provide complementary insights into students' perceived difficulty and class-wide affective dynamics. These findings suggest the pedagogical potential of FER-based metrics to support timely instructional interventions in response to disengagement or confusion. Preliminary results suggest the effectiveness of soft FER. Future work encompasses dataset expansion, multimodal integration, and real-world deployment to enhance both generalizability and practical applicability and may, in the long term, facilitate personalized feedback and adaptive instructional strategies.

Acknowledgement. This work was supported in part by JSPS KAKENHI Grant Numbers JP24K03052, JP22K18006.

References

1. Buono, P., De Carolis, B., D'Errico, F., et al.: Assessing student engagement from facial behavior in on-line learning. Multimedia Tools Appl. **82**, 12859–12877 (2023). https://doi.org/10.1007/s11042-022-14048-8
2. D'Mello, S., Graesser, A.: Autotutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. **2**(4), 1–39 (2012). https://doi.org/10.1145/2395123.2395128
3. Gupta, S., Kumar, P., Tekchandani, R.K.: Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools Appl. **82**, 11365–11394 (2023). https://doi.org/10.1007/s11042-022-13558-9
4. Kok, E.M., Jarodzka, H., Sibbald, M., van Gog, T.: Did you get that? Predicting learners' comprehension of a video lecture from visualizations of their gaze data. Cogn. Sci. **47**(2), e13247 (2023). https://doi.org/10.1111/cogs.13247
5. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2852–2861 (2017). https://doi.org/10.1109/CVPR.2017.277
6. Mahendar, M., Malik, A., Batra, I.: Facial micro-expression modelling-based student learning rate evaluation using VGG-CNN transfer learning model. SN Comput. Sci. **5**(2), 204 (2024). https://doi.org/10.1007/s42979-023-02519-0
7. Mangaroska, K., Sharma, K., Gašević, D., Giannakos, M.: Multimodal learning analytics to inform learning design: lessons learned from computing education. J. Learn. Anal. **7**(3), 79–97 (2020). https://doi.org/10.18608/jla.2020.73.7
8. Pabba, C., Kumar, P.: An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert. Syst. **39**(1), e12839 (2022). https://doi.org/10.1111/exsy.12839
9. Peng, Y., Kikuchi, M., Ozono, T.: Development and experiment of classroom engagement evaluation mechanism during real-time online courses. In: Proceedings of the 24th International Conference on Artificial Intelligence in Education (AIED). Lecture Notes in Artificial Intelligence, vol. 13916, pp. 590–601. Springer (2023). https://doi.org/10.1007/978-3-031-36272-9_48

10. Tang, X., Gong, Y., Xiao, Y., et al.: Facial expression recognition for probing students' emotional engagement in science learning. J. Sci. Educ. Technol. (2024). https://doi.org/10.1007/s10956-024-10143-7
11. Zheng, C., Mendieta, M., Chen, C.: POSTER: a pyramid cross-fusion transformer network for facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3138–3147 (2023). https://doi.org/10.1109/ICCVW60793.2023.00339

Author Index

A
Abou-Khalil, Victoria III-176, III-249
Abu-Rasheed, Hasan III-211
Adamovich, Kseniia II-297
Adkins, Mario I-303
Aguiar, Claire III-60
Akgun, Mahir II-44
Akretche, Idir II-289
Alamri, Ahmed II-328
Aleven, Vincent I-397
Allessio, Danielle II-118
Alpizar-Chacon, Isaac III-275
Alrobai, Amani I-322
Alshamrani, Ghadeer I-476
Aranha, Eduardo H. S. I-415
Araujo Grando, Betsy I-328
Ardimento, Pasquale II-12
Arkün Kocadere, Selay I-47
Armfield, Derek III-111
Arroyo, Ivon I-403, I-409, II-53, II-118
Asswiel, Naif II-289
Azevedo, Roger III-261

B
Babaee, Maryam II-344
Bai, Youheng I-365
Bai, Zhen II-242
Baig, Duaa I-99, I-384
Baker, Ryan S. III-267
Baker, Ryan III-278
Baldwin, Mattew II-297
Balk, Marica I-71
Bang, Hee Jin III-168
Bao, Jiaqi I-269
Barany, Amanda III-278
Barbosa, Carlos Eduardo I-3
Barrett, Alex II-28
Barros, Aristoteles III-160
Belitz, Clara I-291, III-308
Benedetto, Luca III-233
Benton, Laura I-129
Berahmand, Kamal III-77
Bernardi, Mario Luca II-12
Beudt, Susan II-110
Bhattacharya, Soumyadeep III-315
Bhavsar, Arnav III-127
Bhoyar, Rahul Rajkumar II-69
Biswas, Gautam III-218, III-261
Bittencourt, Ig I. I-230
Bittencourt, Ig Ibert I-208, III-160
Bittencourt, Ig I-195
Blanchard, Emmanuel G. II-289, III-241
Blanchard, Nathaniel I-61, II-185, III-144
Borchers, Conrad I-397
Bosarge, Erin II-85
Bosch, Nigel III-308
Botelho, Anthony F. II-266
Botelho, Anthony II-344
Bradford, Mariah III-144
Breitwieser, Jasmin II-274
Brezack, Natalie II-118, II-193, III-119, III-168
Brod, Garvin II-274
Brooks, Christopher III-267
Brusilovsky, Peter III-275
Bulut, Okan III-27
Buzzi, Marina III-43

C
Cabitza, Federico III-201
Cai, Chang I-309
Caines, Andrew III-233
Calabretto, Sylvie I-384
Callahan, Jean-Christophe II-289
Cancela, Pablo III-184
Capdehourat, Germán III-184
Castro, Francisco Enrique Vicente II-53
Cateté, Veronica III-60
Chan, Chun Man II-19
Chan, Wynnie II-193
Chandola, Shrey II-336
Chen, Eason III-111, III-152

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 375–381, 2025.
https://doi.org/10.1007/978-3-031-99264-3

Chen, Gaowei I-297
Chen, Jingxi III-19
Chen, Lujie Karen II-85
Chen, Penghe I-90
Chen, Xiaoran III-103
Chen, Yu I-269
Chepikov, Ivan I-352
Cheung, King Chau II-19
Chiruzzo, Luis III-184
Choe, Ruth I-309
Choi, Jaeyoon III-241
Chounta, Irene-Angelica III-315
Chow, Minyang I-309
Chowdhury, Sancharee Hom II-85
Cimitile, Marta II-12
Clark, Hannah-Beth I-129
Closser, Avery II-344
Cobos, Ruth III-292
Cock, Jade Mai III-261
Cohausz, Lea I-17
Conati, Cristina III-261
Constantin, Aurora I-452
Cordeiro, Thiago I-195, I-208
Costa, Luis Felipe Coimbra I-3
Costley, Jamie II-297
Cristea, Alexandra I. I-322, I-476
Cristea, Alexandra II-328
Cui, Yuan Charles I-439
Cukurova, Mutlu I-391, III-211, III-292

D
da Silva, Alan Pedro I-208
Dąbrowicz-Tlałka, Anita II-282
Dai, Yiling I-119
Damaševičius, Robertas I-47
Daniels, Lia III-86
Daniels, Vijay III-86
Dannath, Jesper I-168
Das, Syaamantak I-181
de Andrade, Henrique I-3
de Witt, Claudia II-69
Dehbozorgi, Nasrin I-433
Demir, Hümeyra III-27
Deriyeva, Alina I-168
Dermeval, Diego I-208, III-160
Di Mitri, Daniele III-292
DiBiano, Robert II-3
Dimitriadis, Yannis II-234, III-225
do Nascimento, João Vitor L. B. I-230
Dong, Wutong I-90
Dowland, Margaux I-129
Du, Junlei II-135
Dueñas, George III-233
Duong-Trung, Nghia II-69

E
Echeverria, Vanessa III-292
Ehara, Yo III-68
El Habib Laraba, Mohamed III-241
England, Matthew I-427
Ercan, Deniz II-344
Esmaeiligoujar, Salah II-344

F
Fan, Xiuyi I-309
Fang, Luyang III-192
Fanguy, Mik II-297
Fechner, Żywilla II-282
Feichtenbeiner, Rolf II-110
Felice, Mariano III-300
Feng, Mingyu II-118, II-193, III-119, III-168
Fernandez Nieto, Gloria III-292
Fernandez, Jose Maria Font I-47
Ferraioli, Valentina III-176, III-249
Ferrato, Alessio I-458
Ferri, Paolo Maria III-225
Firoozi, Tahereh III-86
Fitzgerald, Jack II-185
Flanagan, Brendan I-119, III-249, III-315
Flood, Daniel I-427
Florido, Héctor III-11
Forsyth, Carol II-258
Fujii, Keisuke II-19
Fulantelli, Giovanni III-225

G
Galvan-Sosa, Diana III-233
Gamberini, Carla I-81
Gao, Jie II-61
Gao, Ran II-344
Gaudeau, Gabrielle III-233
Gayne, Will I-129
Geathers, Jadon III-267
Gierl, Mark III-86
Gilbert, Stephen B. I-328
Gong, Na II-85
Góngora, Santiago III-184
Goslen, Alex III-60

Graesser, Arthur C. II-258
Grasman, Raoul II-36
Grawemeyer, Beate I-427, II-321
Gregory, Matthew I-129
Güreş, Fatma Betül III-94

H
Haiman, Aditi I-409
Han, Songhee II-226
Handmann, Uwe III-225
Harsh, Amol II-250
Havinga, Beth III-284
He, Liqun I-391
He, Yutong I-269
Heffernan, Neil T. I-285, I-421
Heffernan, Neil I-340, II-313
Hernández-Leo, Davinia III-11, III-225
Hicke, Yann III-267
Hoel, Tore I-255
Hollenstein, Lena III-176, III-249
Holmes, Wayne III-284
Hosen, Md Biplob II-85
Hou, Xinying II-152
Hu, Xiangen II-258
Hu, Xiaoyong III-103
Hu, Yiling I-269
Huang, Chunwei II-118, III-119, III-168
Huang, Xiaoshan II-61
Huang, Yu-Kai III-152
Hung, Woei II-85
Huynh, Nga II-201

I
Inui, Kentaro II-126
Ipçi, Eylül II-143
Ishioka, Tsunenori I-147
Ismayilova, Aysu I-378
Isotani, Mina I-230
Isotani, Seiji I-195, I-208, I-230, III-160
Iyer, Sridhar I-464

J
Jayatilake, Thumesha I-334
Jeffries, Bryn II-161, II-359
Jeghal, Adil I-242
Jiang, Lan III-308
Jiang, Yang II-258
Jivet, Ioana III-308
Joksimovic, Srecko III-241
Jose, Angelin Mary II-69
Joshi, Omkar I-181
Jung, Changsoo II-185

K
Kaneko, Yuna II-366
Karkalas, Sokratis III-261
Karpov, Ilia I-352
Karunanayake, Ajanie I-482
Käser, Tanja II-143
Kaser, Tanja III-261
Käser, Tanja III-94
Kato, Yoshihiro II-102
Kay, Matthew I-439
Ke, Fengfeng II-28
Kessemtini, Amine II-289
Khan, Imran S. A. II-289
Khosravi, Hassan III-77, III-211
Kikuchi, Masato II-366
Kim, Chaewon II-28, II-226
Kirouchenassamy, Badmavasan III-308
Kizilcec, René F. III-267
Kleiman, Jennifer II-258
Knight, Simon I-482
Kochmar, Ekaterina I-32
Koedinger, Kenneth R. II-152, I-397, III-255
Koedinger, Kenneth III-111, III-152
Konovalov, Vasily III-3
Koprinska, Irena II-161, II-359
Korchak, Anna II-297
Korthals, Luke II-36
Kothiyal, Aditi III-315
Koyejo, Sanmi II-201
Krishnaswamy, Nikhil I-61
Kubota, Shin-Ichiro III-315
Kujawska, Katarzyna II-282
Kunuku, Mourya Teja I-433

L
Laaouina, Loubna I-242
Lai, Joel Weijia I-109
Lai, Yutong I-221
Lallé, Sébastien III-261
Lan, Andrew III-275
Lan, Yunshi III-300
Latif, Ehsan III-135, III-192, III-218
Lavine, Abby II-193
Le, Cuong II-201
Le, Dang II-201

Le, Dong II-201
Le, Tran II-201
Lee, HaeJin III-308
Lee, Morgan P. I-421
Lee, Seiyon M. II-266
Lee, Seiyon II-344
Lee, William I-403
Leporini, Barbara III-43, III-308
Lester, James II-193, III-261
Levine, Jacob II-177
Li, Chenglu II-313
Li, Hai II-313
Li, Hanqi II-152, III-255
Li, Hongming II-266, II-344
Li, Huiyong III-315
Li, Jing III-19
Li, Linlin III-119, III-168
Li, Shuang II-135, III-19
Li, Xueyan II-93
Li, Xueyi I-372
Liao, Guanze III-255
Lim, Fun Siong I-109
Lim, Wen-Chiang Ivan I-285
Lima, Yuri I-3
Lin, Chenyu III-152
Lin, Jionghao II-258, II-352, III-111, III-152
Liu, Joyce Xinle III-27
Liu, Ninghao III-218
Liu, Xueni III-103
Liu, Zifeng II-313
Liu, Zitao I-365, I-372
Lo Duca, Angelica III-43
Loukina, Anastassia III-233
Lu, Haoran III-192
Lu, Yu I-90, I-297, II-77
Luttgen, Kim III-119
Lyu, Bailing II-313

M

Ma, Ping III-192
Maas, Lientje I-71
Macario, Valmir III-160
Majumdar, Rwitajit III-315
Mallavarapu, Aditi II-126
Malloci, Francesca Maridina III-308
Mannan, Sheikh II-185
Marconi, Luca III-201
Marina, Maria III-3
Marques, Leonardo Brandão I-230
Marques, Leonardo I-208
Marquis, Christine III-35
Marras, Mirko III-308
Martin, Kyle I-346
Martínez-Ballesteros, María II-161
Maskeliūnas, Rytis I-47
Mason, Lisa I-61
Matsubayashi, Yuichiroh II-126
Matsuda, Noboru II-126
Maurya, Kaushal Kumar I-32
Mavrikis, Manolis I-391, II-321, III-261
MBaye, Baba I-384
Mbaye, Baba I-99
McGinness, Lachlan I-470
McLaren, Bruce M. I-409
Merchant, Nirav I-137
Min, Wookhee III-60
Minematsu, Tsubasa II-305
Mithun, Paul I-137
Mohammadi, Mehrnoush III-77
Mondada, Francesco III-176, III-249
Monteiro, Mateus III-160
Moore, Steven III-267
Mott, Bradford II-193, III-60
Mu, Su III-103
Mundra, Kushal I-181
Musielak, Magdalena M. II-282
Muszyńska, Karolina I-47

N

Na, Hunhui II-226
Naganuma, Shotaro II-305
Nakamoto, Ryosuke I-119
Nascimento, Luciana Maria Azevedo I-3
Nazaretsky, Tanya II-143
Neto, Raimundo N. B. I-415
Nguyen, Duc II-201
Nguyen, Dung II-201
Nguyen, Huong II-201
Nguyen, Long II-201
Niu, Yazhe II-93
Nixon, Nia III-241
Nóbile, Ana Clara III-184
Noriega-Atala, Enrique I-137
Noviello, Yuri I-446

Nunes, Eduardo Carvalho I-3
Nurbakova, Diana I-99, I-384

O
Ocumpaugh, Jaclyn III-278
Ogata, Hiroaki I-119
Ognibene, Dimitri III-225
Ohno, Aikana II-305
Oka, Haruki I-147
Okada, Kensuke I-147
Omonkulov, Asilbek III-111
Ooge, Jeroen III-211
Oubagine, Ridouane I-242
Ozono, Tadachika II-366

P
Paassen, Benjamin I-168
Paiva, Ranilson I-208
Pal, Sonika I-464
Palansuriya, Charaka I-452
Palomino, Paula T. I-195
Panchenko, Alexander III-3
Pande, Prajakt III-315
Panse, Theresa II-69
Paquette, Luc III-278
Paredes, Jorge I-81
Park, Seehee III-241
Peng, Jingyang II-352
Pian, Yang I-158
Pimentel, Bruno I-230
Pinkwart, Niels II-69
Poell, Tijmen I-71
Poellhuber, Bruno II-210, III-35
Poh, Allison II-53
Portela, Carlos I-195
Porto, Grace Lis I-195
Prasad, Prajish I-464
Prieto, Luis P. II-234
Prince, Brainerd II-250
Psyché, Valéry III-241
Punzo, Veronica III-43

Q
Qi, Xiaofei I-359
Qiu, Wei I-109
Quan, Tho II-201
Queirós, Ricardo I-47

R
Radmehr, Bahar III-94
Rajendran, Ramkumar I-316
Rajmane, Aditya I-316
Ramachandran, Shreyas II-250
Rao, Jiarui II-352
Ravikiran, Manikandan II-169, II-336, III-127
Reichert, Lorenz II-110
Rhaimi, Seyedahmad II-344
Ringstaff, Cathy II-193
Roberts, John I-129
Rodrigues, Luiz I-195, III-160
Rohwer, Kira II-218
Roll, Ido III-261
Roque, Claudia I-81
Rosa, Daniel I-208
Rosenbusch, Hannes II-36
Rotelli, Daniela III-43, III-308
Roy, Normand II-210, III-35
Ruskov, Martin III-225

S
Saba, Janan III-261
Saban, Mohamed II-234
Sadiq, Shazia III-77
Salaj, Ron III-284
Saluja, Rohit II-169, II-336, III-127
Samadi, Mohammad Amin III-241
Santos, Ana Moura I-3
Santos, Olga C. III-315
Sarkar, Soujatya II-169
Scalera, Michele II-12
Schmitt, Lucas II-289
Schneider, Megan II-118
Schroeder, Noah L. II-266
Searle, Emma I-129
Sénécal, Anne-Marie II-210
Seufert, Sabine II-218
Shang, Yi I-359
Sharma, Kshitij I-316
Sharma, Tarun III-127
Shen, Wenyuan II-352
Shi, Yuqian II-53
Shibani, Antonette I-482
Shibukawa, Sachika II-305
Shimmei, Machi II-126
Shindo, Naoki III-52
Shved, Ekaterina III-94
Siddharth, Siddharth II-250
Sikes, Kelsey I-61

Silva, Mariana II-177
Singla, Adish III-94
Skidmore, Edwin I-137
Smith, Andy II-193
Snow, Catherine E. I-230
Sobrinho, Álvaro I-208
Sosnovsky, Sergey III-275
Spikol, Daniel III-292
Sreedharan, Sarath I-61
Srivastava, Namrata III-292
Stamper, John II-152, III-255
Stańdo, Jacek II-282
Stinar, Frank III-308
Stracke, Christian M. III-284
Sukjaitham, Somphop II-274
Sun, Di II-93
Sun, Jingqi III-19
Sun, Zhongtian II-328
Swacha, Jakub I-47

T
Taibi, Davide III-225
Tairi, Hamid I-242
Takami, Kyosuke I-119
Tan, Yoshito I-147
Tang, Xinyi III-111, III-152
Taslimipoor, Shiva III-233
Tenório, Sidarta I-208
Thiessen, Erik III-111
Toda, Armando I-195
Toker, Sacip II-44
Tomaso, Becky II-85
Tran, Phat II-201
Tremblay, Chantal II-210
Troncoso, Alicia II-161
Truong, Sang II-201
Tseng, Ying Jui II-152
Tseng, Ying-Jui III-255
Tudur Sadashiva, Ashwin III-292

U
Uto, Masaki II-126, III-52

V
van Haastrecht, Max I-71
van Rijn, Lars II-69
Vandenberg, Jessica III-60
Venkatesha, Videep III-144
Viscardi, Ignacio III-184
Visser, Ingmar II-36
Vogt, Franziska III-176, III-249
Vogt, Marius III-176, III-249
Vu, Quyen II-201

W
Wakihama, Yukinori II-305
Walkington, Candace II-3
Wang, Chengliang I-221
Wang, Deliang I-297
Wang, Haoming I-221
Wang, Hecong II-242
Wang, Jiayi II-152
Wang, Jingyun I-359, II-328
Wang, Qiao III-300
Wang, Xia II-69
Wang, Ziwei II-161, II-359
Weber, Christian III-211
West, Matt II-177
Wiedbusch, Megan III-261
Wijayasekara, Vihanga Ashinsana I-346
Wiratunga, Nirmalie I-346
Wong, Kester Yew Chong III-292
Wong, Kin Chi II-19
Wong, Tat Wing II-19
Woolf, Beverly II-118
Worden, Eamon I-340
Wrede, Silke Elisabeth II-69
Wu, Bian I-269
Wu, Bo III-267
Wu, Haolun II-61
Wu, Shutong II-242
Wu, Yan III-225

X
Xexéo, Geraldo I-3
Xi, Aprille III-152
Xiao, Ruiwei II-152, III-255
Xiao, Zihan I-221
Xing, Wanli II-313
Xu, Xianlong I-221

Y
Yadav, Deepika I-181
Yang, Dong I-255
Yang, Fumeng I-439
Yeung, Calvin II-19
Yi, Jiyeong I-328
Yin, Yue III-218

Yu, Jeff II-19
Yu, Jinglei II-77
Yu, Xiaoya I-158
Yu, Xinpei III-19
Yuan, Li I-255
Yuan, Zheng III-300

Z
Zamecnik, Andrew III-241
Zapata-Rivera, Diego II-258
Zarestky, Jill I-61
Zarifis, George K. III-225
Zesch, Torsten III-233
Zha, Shenghua II-85
Zhai, Xiaoming II-258, III-135, III-192, III-218
Zhang, Han II-297
Zhang, Jing II-93
Zhang, Kejia I-452
Zhang, Liang II-258
Zhang, Nuodi II-28
Zhang, Shan II-266, II-344
Zhang, Yi I-255
Zhang, Yingbin III-103
Zhang, Yuyin III-103
Zheng, Qinhua II-135
Zhou, Peiyan II-93
Zhou, Qi III-292
Zhou, Yifan III-192
Zhu, Wangda II-313