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

Communications in Computer and Information Science 2592

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 III

Part 3

Springer

AIED PALERMO 2025

Communications in Computer and Information Science 2592

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 III

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 978-3-031-99266-7 ISBN 978-3-031-99267-4 (eBook)
https://doi.org/10.1007/978-3-031-99267-4

© 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 III

Late Breaking Results

Beyond Detection: Rethinking Education in the Age of AI-Writing 3
Maria Marina, Alexander Panchenko, and Vasily Konovalov

Help-Seeking in Problem-Solving: Comparing Generative AI, Peers, and Teacher 11
Héctor Florido and Davinia Hernández-Leo

How Teacher-Provided Need Support Affects Learning Performance in Human-Computer Collaborative Teaching Environments 19
Shuang Li, Jingxi Chen, Jingqi Sun, Jing Li, and Xinpei Yu

Deep Reinforcement Learning for Engagement-Aware Question Selection in Adaptive Assessment Systems 27
Okan Bulut, Joyce Xinle Liu, and Hümeyra Demir

Knowledge, Uses and Perceptions of ChatGPT in Higher Education 35
Christine Marquis, Bruno Poellhuber, and Normand Roy

Inclusive Data Literacy: UDL and AI-Assisted Data Storytelling for BVI Students 43
Marina Buzzi, Barbara Leporini, Angelica Lo Duca, Veronica Punzo, and Daniela Rotelli

Virtual Simulated Patients for Medical Interviews Using Large Language Models with a Self-refinement Mechanism to Suppress Excessive Responses 52
Naoki Shindo and Masaki Uto

Playful Pathways to AI Literacy: Designing NLP Activities for Middle School AI Education 60
Alex Goslen, Claire Aguiar, Jessica Vandenberg, Veronica Cateté, Wookhee Min, and Bradford Mott

Generating Diverse Difficulty Examples in Embedding Vector Spaces via Inverse Embedding 68
Yo Ehara

Knowledge Tracing with A Temporal Hypergraph Memory Network 77
Mehrnoush Mohammadi, Kamal Berahmand, Shazia Sadiq, and Hassan Khosravi

An Augmented Intelligence System for Automated Quality Control and Feedback Generation of Multiple Choice Test Items 86
Tahereh Firoozi, Lia Daniels, Vijay Daniels, and Mark Gierl

ClickSight: Interpreting Student Clickstreams to Reveal Insights on Learning Strategies via LLMs 94
Bahar Radmehr, Ekaterina Shved, Fatma Betül Güreş, Adish Singla, and Tanja Käser

Effects of Human and Automated Feedback on Pre-service Teachers' Nonverbal Teaching Skills During Microteaching 103
Su Mu, Xiaoran Chen, Xueni Liu, Yuyin Zhang, Yingbin Zhang, and Xiaoyong Hu

Avalon: A Human-in-the-Loop LLM Grading System with Instructor Calibration and Student Self-assessment 111
Derek Armfield, Eason Chen, Asilbek Omonkulov, Xinyi Tang, Jionghao Lin, Erik Thiessen, and Kenneth Koedinger

Empowering Teachers with Technology: A National Study on a Formative Assessment Platform 119
Mingyu Feng, Linlin Li, Chunwei Huang, Natalie Brezack, and Kim Luttgen

CASSA: Context-Aware Self-attention with Global Context Suppression and Relevance Modulation for MCQ Difficulty Estimation 127
Manikandan Ravikiran, Tarun Sharma, Arnav Bhavsar, and Rohit Saluja

Privacy-Preserved Automated Scoring Using Federated Learning for Educational Research 135
Ehsan Latif and Xiaoming Zhai

Dude, Where's My Utterance? Evaluating the Effects of Automatic Segmentation and Transcription on CPS Detection 144
Videep Venkatesha, Mariah Bradford, and Nathaniel Blanchard

VTutor: An Animated Pedagogical Agent SDK that Provide Real Time Multi-model Feedback 152
Eason Chen, Chenyu Lin, Yu-Kai Huang, Xinyi Tang, Aprille Xi, Jionghao Lin, and Kenneth Koedinger

"Small Device, Big Decision:" Comparing Lightweight LLMs' Computational Performance and Output Quality for AIED Unplugged 160
Mateus Monteiro, Aristoteles Barros, Luiz Rodrigues, Diego Dermeval, Ig Ibert Bittencourt, Seiji Isotani, and Valmir Macario

Personalized Math Learning for Young Learners: Preliminary Findings from a Rigorous Study of a Game-Based Program 168
Mingyu Feng, Linlin Li, Chunwei Huang, Natalie Brezack, and Hee Jin Bang

Teachers' Perspectives on Using and Teaching Artificial Intelligence in Early Primary Education 176
Marius Vogt, Valentina Ferraioli, Victoria Abou-Khalil, Lena Hollenstein, Francesco Mondada, and Franziska Vogt

Automatic Assessment of Spoken English for Uruguayan Spanish L1 Young Learners 184
Ana Clara Nóbile, Ignacio Viscardi, Pablo Cancela, Germán Capdehourat, Luis Chiruzzo, and Santiago Góngora

Efficient Multi-task Inferencing: Model Merging with Gromov-Wasserstein Feature Alignment 192
Luyang Fang, Ehsan Latif, Haoran Lu, Yifan Zhou, Ping Ma, and Xiaoming Zhai

Tutor or Mentor. A Comparative Usability Study of AI Roles in Higher Education 201
Luca Marconi and Federico Cabitza

Workshops

XAI-Ed: Half-Day Workshop on Pedagogy-Founded Explainable AI for Transparent, User-Centered AI in Education 211
Hasan Abu-Rasheed, Mutlu Cukurova, Hassan Khosravi, Jeroen Ooge, and Christian Weber

Workshop on Epistemics and Decision-Making in AI-Supported Education 218
Ehsan Latif, Ninghao Liu, Gautam Biswas, Yue Yin, and Xiaoming Zhai

D-SAIL - Transformative Curriculum Design: Digitalisation, Sustainability, and AI Literacy for 21st Century Learning 225
Martin Ruskov, Dimitri Ognibene, Davinia Hernández-Leo, Davide Taibi, Yannis Dimitriadis, Giovanni Fulantelli, Uwe Handmann, George K. Zarifis, Yan Wu, and Paolo Maria Ferri

Second Workshop on the Automatic Evaluation of Learning and Assessment Content 233
Luca Benedetto, Diana Galvan-Sosa, George Dueñas, Shiva Taslimipoor, Gabrielle Gaudeau, Andrew Caines, Anastassia Loukina, and Torsten Zesch

Human-AI Collaboration and Culture For Equity and Inclusion (Half-Day Workshop) 241
Nia Nixon, Srecko Joksimovic, Andrew Zamecnik, Mohammad Amin Samadi, Jaeyoon Choi, Seehee Park, Valéry Psyché, Mohamed El Habib Laraba, and Emmanuel G. Blanchard

Preparing Teachers for an Equitable AI Integration in Education 249
Victoria Abou-Khalil, Lena Hollenstein, Valentina Ferraioli, Marius Vogt, Brendan Flanagan, Franziska Vogt, and Francesco Mondada

AI Literacy For All: 1st International Workshop on AI Literacy Education For All 255
Ruiwei Xiao, Ying-Jui Tseng, Hanqi Li, Guanze Liao, John Stamper, and Kenneth R. Koedinger

Workshop on Advancements in AI-Supported Exploratory Learning (AI4EL) 261
Manolis Mavrikis, Sébastien Lallé, Roger Azevedo, Gautam Biswas, Jade Mai Cock, Cristina Conati, Sokratis Karkalas, Tanja Kaser, James Lester, Ido Roll, Janan Saba, and Megan Wiedbusch

Applications of Generative AI to Support Teaching and Learning in Higher Education: A Half-Day Workshop 267
René F. Kizilcec, Ryan S. Baker, Christopher Brooks, Jadon Geathers, Yann Hicke, Steven Moore, and Bo Wu

Intelligent Textbooks: The Sixth International Workshop 275
Sergey Sosnovsky, Peter Brusilovsky, Andrew Lan, and Isaac Alpizar-Chacon

Full-Day Tutorial: Conducting in-the-Moment Data-Driven Classroom Interviews (DDCI) During Digital Learning Using Quick Red Fox (QRF) 278
Jaclyn Ocumpaugh, Luc Paquette, Amanda Barany, and Ryan Baker

Ethical Artificial Intelligence and Education: The Need for International Regulation to Foster Human Rights, Democracy and Equity 284
Christian M. Stracke, Beth Havinga, Wayne Holmes, and Ron Salaj

The Second International Workshop on Multimodal Artificial Intelligence in Education (MAIEd'25) 292
Daniele Di Mitri, Namrata Srivastava, Gloria Fernandez Nieto, Vanessa Echeverria, Ruth Cobos, Ashwin Tudur Sadashiva, Daniel Spikol, Kester Yew Chong Wong, Qi Zhou, and Mutlu Cukurova

NLP and Generative AI for Language Learning and Assessment: Synergies Between Research and Practice 300
Zheng Yuan, Mariano Felice, Yunshi Lan, and Qiao Wang

Fair4AIED 2025: First International Workshop on Fairness in Algorithmic Decision-Making for Education 308
Clara Belitz, Nigel Bosch, Lan Jiang, Ioana Jivet, Badmavasan Kirouchenassamy, HaeJin Lee, Barbara Leporini, Francesca Maridina Malloci, Mirko Marras, Daniela Rotelli, and Frank Stinar

Multimodal Interactions and Explainable AI for Reflective Physical and Online Learning: MiXai^learn Workshop 315
Rwitajit Majumdar, Huiyong Li, Brendan Flanagan, Shin-Ichiro Kubota, Soumyadeep Bhattacharya, Aditi Kothiyal, Prajakt Pande, Olga C. Santos, and Irene-Angelica Chounta

Author Index 321

Late Breaking Results



Beyond Detection: Rethinking Education in the Age of AI-Writing

Maria Marina[1,2(✉)], Alexander Panchenko[1,2], and Vasily Konovalov[1,2,3]

[1] AIRI, Moscow, Russia
{marina,konovalov}@airi.net
[2] Skoltech, Moscow, Russia
[3] Moscow Institute of Physics and Technology, Moscow, Russia

Abstract. As generative AI tools like ChatGPT enter classrooms, workplaces and everyday thinking, writing is at risk of becoming a formality – outsourced, automated and stripped of its cognitive value. But writing is not just output; it is how we learn to think. This paper explores what is lost when we let machines write for us, drawing on cognitive psychology, educational theory and real classroom practices. We argue that the process of writing – messy, slow, often frustrating – is where a human deep learning happens. The paper also explores the current possibilities of AI-text detection, how educators can adapt through smarter pedagogy rather than bans, and why the ability to recognize machine-generated language may become a critical literacy of the 21st century. In a world where writing can be faked, learning cannot.

Keywords: AI-Assisted Writing · AI-text detection · Pedagogy

1 Introduction

"I write because I don't know what I think until I read what I say."
—Flannery O'Connor

We are now seeing a growing trend where various roles are outsourcing their work to AI. Administrators are delegating condolence emails to automated tools, job applicants are using AI to craft tailored cover letters, marketers are generating entire ad campaigns with a few prompts, students are turning to AI to write their five-paragraph essays, and professors are relying on AI to grade those very same papers. "When machines are grading machines, one should ask: what have we lost?" [10].

Digital technologies, while offering unprecedented convenience, can also erode essential cognitive skills such as learning and memory retention [2]. When AI tools generate text on our behalf, they effectively short-circuit the deep thinking involved in the writing process. Writing is not just about producing words;

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 3–10, 2025.
https://doi.org/10.1007/978-3-031-99267-4_1

it is a cognitive exercise that involves refining ideas, questioning initial assumptions, and often rewriting or restructuring entirely. This reflection fosters critical thinking and deeper understanding [3].

When it comes to memory, recent research highlights a growing trend of relying on digital tools to store and retrieve information, an effect known as cognitive offloading. People are increasingly turning to the Internet as an external memory system, decreasing their reliance on internal recall [30]. This shift, while convenient, may come at the cost of weakening our natural memory functions.

Adding to this, individuals who searched online for information not only remembered less but also developed an inflated sense of their own knowledge [6]. In this study, participants who had looked up information on one topic (let us say topic X) were more likely to believe that they were knowledgeable about a completely unrelated topic (topic Y), simply because they felt confident that they could find the answer if needed. This illusion of understanding, fueled by the accessibility of online information, creates a false sense of expertise and undermines true learning. In contrast, while writing, the information is better memorized [26].

These findings align with broader concerns about how digital convenience shapes our cognitive habits. As psychologist Betsy Sparrow explains in her work on the "Google Effect", people are increasingly remembering how to find information rather than remembering the information itself [27]. This raises important questions about what happens to deep knowledge when we treat search engines as mental extensions rather than supplements.

2 Why Do We Still Need to Teach Writing Skills?

In an era increasingly defined by automation and artificial intelligence, the role of writing in education is undergoing a profound shift. Rather than serving only as preparation for a career or a professional skillset, writing is being reframed – particularly in some forward-thinking business schools – as a foundational tool for living a thoughtful, intellectually engaged life [25]. This liberal arts-inspired approach suggests that the true value of writing lies not in producing polished prose, but in what the act of writing enables: deliberate thinking, deeper reflection, and a clearer understanding of one's own ideas.

This distinction is especially evident in the classroom, where students often fall into two broad camps. Some focus primarily on efficiency and outcomes – grades, degrees – and view writing as a hurdle to be cleared as quickly as possible. For them, generative AI tools are shortcuts that enable faster completion of assignments with minimal cognitive effort. But others see writing differently. These students treat it as a tool for discovery – a means to figure out what they truly think and how best to express it. The process is not always easy. It involves confronting vague ideas, reshaping them through drafts, and wrestling with language to articulate precise, persuasive arguments. Yet, in doing so, they develop sharper, more nuanced reasoning skills and a deeper intellectual maturity [10].

As Justin E.H. Smith points out, anti-intellectualism has long been a part of our cultural landscape, and the rise of AI only makes the struggle against it

more difficult. A core purpose of education is to push back against this tendency – to help students experience the value of learning from the inside rather than trying to convince them from the outside. Some students come eager to engage, but many are naturally resistant to learning efforts, especially when those efforts demand deep thinking. Still, as they work through their assignments, something often changes. They begin to sense the difference between uninformed opinions and carefully developed thoughts. Realization that clear, informed thinking feels different and has power is a pivotal moment in their intellectual growth. The concern with generative AI is that it can completely disrupt this process, offering a shortcut that skips over the essential struggle of learning how to think well [32].

Writing is the very act through which people learn to think clearly [21]. Until one sits down to write, it is difficult to fully grasp the logic or coherence of one's ideas. The process slows us down, forcing thoughts onto the page where they can be seen, questioned, and improved. Scholars such as Hays [9], Lawrence [15], and Werse [33] reinforce this idea: writing is an iterative activity that deepens understanding by pushing us to organize and reorganize our thoughts for a particular audience. When that process is outsourced to AI, students may complete the task, but they forfeit the deeper cognitive and developmental benefits that come with struggling through it themselves. Naomi S. Baron states that relying on AI to generate content "diminishes opportunities to think out problems for ourselves". She also mentioned the risk of a slippery slope, where students start letting generative AI control the content and style of their writing, leading to a diminished sense of ownership [34].

3 Impossibility of Perfect Detection of AI

There are currently four main research directions aimed at addressing the challenge of AI-generated text detection. The first treats the task as a binary classification problem, using neural network-based detectors [20,31]. The second explores zero-shot detection methods [18], relying on per-token log probabilities and thresholding to identify AI-generated text. The third line of work focuses on watermarking, which facilitates detection by embedding identifiable patterns into generated text. For example, a soft watermarking technique that categorizes tokens into "green" and "red" lists [11]. A watermarked LLM is biased to select tokens from the green list, which is generated pseudo-randomly based on the prefix tokens. A detector can then classify a passage as AI-generated if it contains a disproportionately high number of green-list tokens. Finally, information retrieval-based methods [13] detect AI-generated content by searching a database for semantically similar matches.

However, recent studies have shown that all of these approaches have significant limitations. [24] demonstrate that five rounds of recursive paraphrasing reduce detection accuracy on watermarked text to below 20%. Similarly, the AUROC score of DetectGPT [18] drops from 96.5% to 59.8% after such attacks. Furthermore, the authors show that for any two distributions, where H is human text and M is AI-generated text – the best possible detector's

AUROC score decreases as the total variation distance TV(M, H) between them shrinks. Another study finds that, with the exception of Pangram [14] (which is closed-source), all existing detection systems perform poorly under evasion strategies [23].

As these findings show, there is currently no reliable, open-source solution for detecting AI-generated text. As human and AI-generated outputs grow more alike – for instance, when TV < 0.2, – even the best possible detectors yield unreliable performance (AUROC <0.7) [24].

Moreover, even if watermarking were effective, it introduces several additional concerns. **Trade-offs in quality:** Watermarks may reduce fluency and quality, hurting user experience. Balancing imperceptibility and text quality remains a major challenge [19]. **Lack of standardization:** There is no widely accepted standard for watermarking AI-generated text. Different organizations developing their own methods could lead to compatibility issues and inconsistent detection outcomes across platforms [28]. **Ethical and privacy concerns.** Mandating watermarking could raise serious ethical and privacy issues, such as infringing on free expression or enabling surveillance. Additionally, watermarking may fail to address deeper problems like plagiarism or misinformation [16].

4 Use Cases in Secondary Schools

As technology continues to evolve, educators find themselves navigating a rapidly changing landscape. One striking parallel is the introduction of generative AI tools like ChatGPT in writing classrooms. In the same way that calculators were once met with skepticism, but are now widely accepted as a necessary tool in math education, AI is becoming an integral part of teaching writing. But like the gym, where results are gained not from watching others but from doing the work yourself, the real value lies in the student's engagement with the process. It's not about AI writing for students, but about using it to enhance their own thinking and writing skills.

All best practices used by both school teachers and university professors to teach writing are divided into three categories.

How to Use AI For Teaching Writing Effectively. AI can serve as an assistant, not a replacement. It can help students in early stages, such as organizing their thoughts and generating ideas.

- *AI-Assisted Outlining.* Cherie Shields, a high school teacher in Oregon, asked her students to use ChatGPT to generate outlines for essays comparing two 19th-century short stories. The students then wrote their essays by hand, deepening their understanding and refining their ideas. This strategy helped students not only interact with AI but also understand how to guide it to generate useful responses [22].
- *Enhancing Visual Imagination.* Matt Cave, a head teacher at Willowdown, encourages students to feed their descriptive writing into AI to generate images. This helps them visualize what their words convey and compare it

with their peers' interpretations: "they can then discuss with their classmates whether that was the image they expected to be in the reader's head." [1]
- *Reflection and Revision.* One strategy involves asking students to submit two versions of their work: the original (non-AI), and the corrected (post-AI) version. In the post-AI version students have to comment on what mistakes AI has made and what the student is not sure about. This nice exercise teaches them to draw a line between legitimate and illegitimate use of resources [29].

Harder For The Students to Cheat Than Not to Cheat. Instructors are developing clever ways to reduce the temptation to rely on generative AI – not by banning it, but by designing assignments that are actually easier to complete honestly than to fake.

- *Focused Mini-Assignments Over Full Papers.* Rather than assigning broad, full-length essays, some educators ask students to complete short, tightly focused writing tasks connected to key skills. For instance, students might be prompted to write a paragraph of lively prose, analyze a real-life observation, or turn a personal moment into a more universal insight. These concise tasks are harder to delegate to AI and often resonate more with students, making them more inclined to do the work themselves [4].
- *Make Instructions Too Specific to Outsource.* Some instructors carefully craft prompts with detailed structure and formatting expectations that are tough to replicate with a simple AI query. For example, assignments may also be broken into weekly milestones to ensure consistency across drafts, making it much harder to outsource work submitted a week ago [29].
- *Constrained Vocabulary Challenge.* An effective vocabulary – building activity involves giving students a list of 20 words and requiring them to use exactly 10 of them – each only once in a short writing piece [29].

Other Ideas For Encouraging Authentic Writing. Alongside structured prompts and clever assignment design, instructors are also experimenting with broader strategies to promote originality and student voice.

- *In-Class Free-writing to Capture Voice.* Having students write short, unassisted paragraphs in class provides a snapshot of their natural writing voice. Over time, these samples help instructors track each student's style and recognize when later submissions deviate too far from it. This strategy not only discourages academic dishonesty but also gives students a chance to reflect on their developing voice [17].
- *Personal Progress Grading.* Instead of comparing students to each other, some instructors evaluate work based on individual progress. By gathering an initial writing sample early in the term, teachers can measure each student's growth over time. This approach eases anxiety, encourages effort, and allows students to take pride in their own development, and making shortcuts like AI-generated writing less appealing [29].
- *Assign What AI Doesn't Know.* One effective way to limit AI's usefulness in assignments is by asking students to write about recent news stories that fall

outside AI training data. For example, students might be asked to analyze a current event that happened after a known AI model's knowledge cutoff. Alternatively, they can exploit LLMs refusal to respond to politics-related queries. In one case, a student used AI to translate Burna Boy's Nigerian Pidgin but it was necessary to rely on online forums for interpreting the political messages of the song [8].

It is also important to note that banning access to generative technologies is counterproductive. Students who are determined to use these tools will likely find a way, but such restrictions tend to favor those whose first language is English, who can afford access to paid large language models or applications, and who are more digitally literate. Thus, increasing the gap between less and more priviliged students even more [7].

5 Recognizing AI-Generated Texts as a Critical Skill

Recent studies estimate that over 5% of recently-published Wikipedia articles [5] and 10% of PubMed abstracts published in 2024 [12] were written by AI. This highlights the necessity to develop skills for detecting AI-generated content.

While the theoretical section demonstrates that evasion techniques can render AI-generated texts nearly undetectable, recent research [23] offers a more optimistic view for human detection. It shows that individuals who regularly use LLMs for writing tasks are significantly better at identifying AI-generated texts. In a study involving five such experienced annotators, only 1 out of 300 nonfiction articles was misclassified when decisions were made by majority vote.

Qualitative analysis revealed that these experts relied not only on lexical cues (such as "AI vocabulary") but also on more nuanced elements like clarity, originality, and formality – features that are often difficult for automatic detectors to assess. This expertise proved robust across various models (including GPT-4o and Claude 3.5 Sonnet) and was unaffected by common evasion strategies, such as paraphrasing and humanization (i.e., prompts crafted using expert-derived instructions to simulate human writing). Remarkably, their detection capabilities extended even to content generated by the newly released o1-pro model.

If humans can be trained to recognize AI-generated texts, then this skill must become a core competency for both educators and students. A compelling real-world example was documented by Craig Meyer [17]:

"The perfect sentences and perfect grammar. Finally, it dawned on me. Students don't write like this; I don't even write like this. It was too perfect. Yet there was something else, something much more important. It was dead. Lifeless. Perhaps even boring. Human writers tend to inject anecdotes, humor, slight digressions, stylistic preferences, and other perceptible creations of our genius. This paper had none of those, and once I saw it, I couldn't unsee it. "

6 Conclusion

The rise of generative AI confronts educators with a paradox: while these tools promise efficiency, they risk losing the very process that makes education trans-

formative. When students bypass the struggle of writing, they forfeit the deep reflection that sharpens thinking, fosters memory, and stimulates intellectual independence. Instead of banning AI, we must redesign teaching to reward authentic voice, progress over perfection, and make cheating harder than learning. If we want students to write like humans, we must show them that writing *makes* them human. And in a future where synthetic text is everywhere, the ability to think clearly, write originally, and detect what's real won't just be valuable – it will be essential.

P.S.: For writing of this article, AI assistant tools were only used for grammar checking and paraphrasing of several individual sentences.

References

1. Adams, R.: The English schools looking to dispel 'doom and gloom' around AI (2025). https://www.theguardian.com/education/2025/mar/06/the-english-schools-looking-to-dispel-doom-and-gloom-around-ai Accessed 5 Apr2025
2. Baron, N.S.: Know what? how digital technologies undermine learning and remembering. J. Pragmat. **175**, 27–37 (2021)
3. Baron, N.S.: How chatgpt robs students of motivation to write and think for themselves (2023). https://theconversation.com/how-chatgpt-robs-students-of-motivation-to-write-and-think-for-themselves-197875 Accessed 5 Apr 2025
4. Bogost, I.: AI cheating is getting worse (2024). https://www.theatlantic.com/technology/archive/2024/08/another-year-ai-college-cheating/679502/ Accessed 5 Apr 2025
5. Brooks, C., Eggert, S., Peskoff, D.: The rise of AI-generated content in wikipedia (2024). https://arxiv.org/abs/2410.08044
6. Fisher, M., Goddu, M.K., Keil, F.C.: Searching for explanations: how the internet inflates estimates of internal knowledge. J. Exp. Psychol. Gen. **144**(3), 674 (2015)
7. Furze, L.: AI detection in education is a dead end (2024). https://leonfurze.com/2024/04/09/ai-detection-in-education-is-a-dead-end/ Accessed 5 Apr 2025
8. Girdharry, K., Khachatryan, D.: Meaningful writing in the age of generative artificial intelligence. Double Helix **11**(1) (2023)
9. Hays, J.N., et al.: The Writer's Mind: Writing as a Mode of Thinking. ERIC (1983)
10. Jocelyn, C.: Teaching writing in the age of llms. https://chimesnewspaper.com/53967/opinions/teaching-writing-in-the-age-of-llms/ (2024)
11. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A watermark for large language models. In: International Conference on Machine Learning, ICML. vol. 202, pp. 17061–17084 (2023)
12. Kobak, D., González-Márquez, R., Ágnes Horvát, E., Lause, J.: Delving into chatgpt usage in academic writing through excess vocabulary (2025). https://arxiv.org/abs/2406.07016
13. Krishna, K., Song, Y., Karpinska, M., Wieting, J., Iyyer, M.: Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems (2023)
14. Labs, P.: Industry leading models to identify ai generated text (2024). https://www.pangram.com/our-model/ai-detection Accessed 6 Apr 2025
15. Lawrence, M.S., et al.: Writing as a thinking process. University of Michigan Press (1972)

16. Lisinska, J.: Why watermarking text fails to stop misinformation and plagiarism (2024). https://itif.org/publications/2024/09/18/why-watermarking-text-fails-to-stop-misinformation-and-plagiarism/ Accessed 5 Apr 2025
17. Meyer, C.A.: Generative AI: the voice of the other. Double Helix **11**(1), 1–5 (2023)
18. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: Detectgpt: zero-shot machine-generated text detection using probability curvature. In: International Conference on Machine Learning, ICML. vol. 202, pp. 24950–24962 (2023)
19. Nature: AI watermarking must be watertight to be effective. Nature **634**(8046), 753 (2024). https://doi.org/10.1038/d41586-024-03418-x. https://www.nature.com/articles/d41586-024-03418-x Accessed 5 Apr2025
20. OpenAI: Gpt-2: 1.5b release (2019). https://openai.com/research/gpt-2-1-5b-release Accessed 6 Apr 2025
21. Pugh, D.S.: How to get a PhD: a handbook for students and their supervisors. Open University Press (2014)
22. Roose, K.: Don't ban chatgpt in schools. teach with it. International New York Times pp. NA–NA (2023)
23. Russell, J., Karpinska, M., Iyyer, M.: People who frequently use chatgpt for writing tasks are accurate and robust detectors of AI-generated text (2025). https://arxiv.org/abs/2501.15654
24. Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can AI-generated text be reliably detected? (2025). https://arxiv.org/abs/2303.11156
25. Schatten, J.: Will artificial intelligence kill college writing? The Chronicle of Higher Education (2022)
26. Silva, A.M., Limongi, R.: Writing to learn increases long-term memory consolidation: a mental-chronometry and computational-modeling study of "epistemic writing". J. Writ. Res. **11**(1), 211–243 (2019)
27. Sparrow, B., Liu, J., Wegner, D.M.: Google effects on memory: cognitive consequences of having information at our fingertips. science **333**(6043), 776–778 (2011)
28. Srinivasan, S.: Detecting AI fingerprints: a guide to watermarking and beyond. https://www.brookings.edu/articles/detecting-AI-fingerprints-a-guide-to-watermarking-and-beyond/ (2024). Accessed 5 Apr 2025
29. StackExchange User: Is there a way to stop students using AI to generate essays? https://academia.stackexchange.com/questions/208862/is-there-a-way-to-stop-students-using-ai-to-generate-essays (2024)
30. Storm, B.C., Stone, S.M., Benjamin, A.S.: Using the internet to access information inflates future use of the internet to access other information. Memory **25**(6), 717–723 (2017)
31. Voznyuk, A., Konovalov, V.: DeepPavlov at SemEval-2024 task 8: Leveraging transfer learning for detecting boundaries of machine-generated texts. In: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). pp. 1821–1829 (Jun 2024).https://doi.org/10.18653/v1/2024.semeval-1.257
32. Weinberg, J.: A philosopher's reflections on teaching in a world with AI (2025). https://dailynous.com/2025/03/07/a-philosophers-reflections-on-teaching-in-a-world-with-ai/Accessed 5 Apr 2025
33. Werse, N.R.: What will be lost? critical reflections of chatgpt, artificial intelligence, and the value of writing instruction. Double Helix **11** (2023)
34. Wikipedia contributors: Chatgpt in education — wikipedia, the free encyclopedia (2024). https://en.wikipedia.org/wiki/ChatGPT_in_education Accessed 5 Apr 2025

Help-Seeking in Problem-Solving: Comparing Generative AI, Peers, and Teacher

Héctor Florido(✉) and Davinia Hernández-Leo(✉)

Universitat Pompeu Fabra, Barcelona, Spain
hector.florido@estudiant.upf.edu, davinia.hernandez-leo@upf.edu

Abstract. This study explores help-seeking supported by Generative Artificial Intelligence (GenAI), peers and the teacher during problem-solving sessions. The research aims to collect and analyze data from various sessions in a vocational training setting to compare which are the most effective doubt-resolution strategies for tasks related to software programming. By involving students in the analysis, the study compares their perceptions of the efficacy of responses from GenAI tools, peers, and teachers against their initial expectations. Results indicate that students unexpectedly rate the GenAI tool's responses more favorably than those from their peers and the teacher. This paper discusses these findings, informing the opportunities for effective integration of GenAI tools in educational settings depending on the context. The comparative analysis seeks to illuminate the complementary roles of GenAI tools, peers, and teachers in the design of human-machine co-orchestration strategies involving help-seeking in problem-solving classrooms.

Keywords: Help-seeking · classroom orchestration · generative AI · problem solving

1 Introduction

A problem-solving session in a vocational training classroom aims to let the students practice the concepts explained in recital sessions facing different situations. Research on help-seeking behavior has been widely conducted in online and hybrid learning environments, where students often face challenges due to limited personal interaction. Obtaining support when communicating a problem or difficulty is the expression of help-seeking [1]. Students seek assistance from teachers and other participants during recital and problem-solving sessions. Aleven [2] explores the complex dynamics of help-seeking behavior in interactive learning environments. They analyze how factors such as prior knowledge, metacognitive skills, cognitive development, gender, goal orientation, and epistemological beliefs influence both the likelihood and effectiveness of help-seeking.

Help-seeking is often a fundamental aspect of the learning process, providing students with opportunities to develop strategies to overcome difficulties, clarify concepts, and enhance their understanding. As part of their tasks, teachers need to orchestrate

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 11–18, 2025.
https://doi.org/10.1007/978-3-031-99267-4_2

learning scenarios considering the potential help-seeking needs and finding the best approaches to manage them effectively and efficiently. However, the burden on teachers to ensure successful orchestration has been found to be challenging [3]. By fostering inclusive and supportive learning spaces, classroom orchestration technology [4] offers the potential to facilitate this management in ways that optimize how students seek assistance confidently and receive prompt, precise responses to their questions in the classroom. The advancements in artificial intelligence (AI) have significantly modified various aspects of our lives. The education sector, in particular, is being influenced by these innovations, especially with the advent of Generative AI (GenAI). Approaches to the integration of AI in education have generated diverse perspectives, challenges, and opportunities [5]. GenAI has the potential to evolve educational technology and redefine classroom orchestration [6].

GenAI has become a remarkable advancement in artificial intelligence due to its implications in different societal domains, as it is able to generate new content such as text, images, and music by identifying and learning from existing data patterns, among other AI techniques. Unlike other AI systems that focus on classification or analysis, GenAI models use sophisticated techniques to generate new outputs. For example, a GenAI model trained on a large text dataset (i.e. a Large Language Model, LLM) can produce coherent and contextually relevant written content. This technology offers potential for innovation in fields like creative writing, digital art, and beyond [7, 8] and in diverse support areas within education [6, 9].

AI technologies, including chatbots, robots, and automated assessments, have significantly influenced the four key educational domains: learning, teaching, assessment, and administration [10]. These advancements have transformed traditional educational practices by enabling personalized learning experiences, improving teaching strategies, streamlining assessment procedures, and enhancing administrative efficiency [11]. However, AI cannot fully replace the pedagogical expertise of teachers, as it lacks the deep understanding needed to tailor teaching methods to individual students and considering the context-specific needs and resources. Moreover, ethical concerns, particularly regarding data privacy, bias, reliability as well as accountability, pose considerable challenges to AI's integration in education [12, 13].

The success of AI in education relies on effective collaboration between humans and AI. When educators leverage AI's capabilities alongside their own expertise, they can create personalized, engaging, and equitable learning experiences. As Cukurova [12] and Järvelä [14] highlight, this partnership fosters a more effective and ethical educational environment. AI can assist teachers by automating routine tasks, analyzing student performance, and providing personalized feedback, while educators remain essential in overseeing the generated recommendations, guiding learning and social development. Together, humans and AI can offer enhanced levels of support that improve efficiency and effectiveness in both teaching and learning.

The design of GenAI classroom orchestration systems supporting help-seeking in problem-solving scenarios should consider the mentioned opportunities, from the perspective of human-AI collaboration to maximize the potential benefits while minimizing the risks and ethical implications. From this perspective, in this paper we explore the potential of incorporating a GenAI feature in the help-seeking orchestration of software

programming problem-solving classrooms. In particular, the paper seeks to answer the following research questions: RQ1) How does it compare students' assessment of the help received from teachers, peers or GenAI tools? RQ2) What are the students' preferences regarding their expectations for help from teachers, peers, or GenAI tools? RQ3) Are students aware of the importance of the prompts when obtaining results from a GenAI tool help seeking?

2 Methodology

For the study, 40 participants were involved in the problem-solving sessions involved in the study, all of them over 18 years old. The sessions were developed in a vocational training high school in Barcelona, in a course on software programming, more concrete in Java language programming [15]. Only one teacher was involved in the sessions and was the person who introduced the students to a new concept of the subject and guided them through the problem-solving sessions.

Four sessions were planned in this study because the analysis of the data between sessions was very important for preparing the next step. The same group of students and the same teacher participated in all four sessions.

The first session was focused on explaining the concepts of the subject in a way that encouraged students to ask questions. The entire session was done without the opportunity for students to interact with any Gen AI or even search on Google for help. After the theoretical part was completed, the students were given some exercises related to the subject [15], and they had to solve them without help, using the programming tools commonly used in their classes. When they finished, the questions were collected via a questionnaire for further analysis.

Repeated and nonsense questions were removed for the next steps. Multiple-in-one questions were also split into different questions in order to obtain more concrete results in the following sessions. A questionnaire was generated with all the questions submitted by the students and randomized in order to assign different questions to each student involved in the study.

The second session was planned to have three different parts. First, the students had to answer the selected questions from the form. These responses were generated based on the knowledge of every student without GenAI access and without using Google to search for help. After all the questions were answered, the second part of the session required them to answer the same questions, but this time with GenAI help. From this interaction, all the markdown code generated by the GenAI was requested, along with the prompt each student used to obtain the answer to the questions.

At this point, two different responses for each question had been collected from peers, that include their own response and a ChatGPT response, in addition to the ones generated by the teacher, including his own response and the ones generated using GenAI tools. The tools chosen for this specific experimental session were, first of all, ChatGPT with prompts from both students and the teacher, Gemini, only using the same prompt as ChatGPT written by the teacher, and DeepSeek, using the same prompt as well. For every question, six different responses were collected.

The third session involved the distribution to students of a questionnaire with three of the documents containing the six different responses [15] in order for them to read all

the answers and rate them from 1 to 6 points each, where 1 point meant the least efficient, structured, or synthesized answer, and 6 points meant the best one. The factors taken into account were efficiency, accuracy, and clarity of examples or explanations in order to determine whether the answers were more or less useful in resolving the question. These scores were summarized and stored in a graph in order to present them in the next and last experimental session with the student group.

For the last session, graphs with the results obtained from the previous sessions were prepared, along with some follow-up questions needed to understand these results together with the teacher and students involved in the experimental sessions. The students were asked to classify, based on their opinion, the results obtained in the previous sessions before seeing the actual results. After this, the results were revealed to the students and discussed with the teachers, exploring the meaning of all this data, followed by a questionnaire where they could express their opinion on the obtained results. This last session ended with a debate to gather students' opinions on when they should use GenAI in the classroom and whether any of them preferred to skip these types of tools in order to acquire knowledge on their own.

3 Findings

3.1 Student's Expectations and Assessments of the Help Received

The data obtained from the last questionnaire was grouped into the different categories of responses given by the teacher, peers, ChatGPT, Gemini, and DeepSeek. All the scores assigned were summed and divided by the total number of scores in order to obtain the percentage across the whole ranking. Table 1 **(column 1–2)** shows the results obtained from this questionnaire and how similar all the final scores were.

Table 1. Scores obtained from students' assessment ratings (**columns 1–2**) and previous rating expectations from students (**columns 3–4**).

Group	Score	Percentage	**Expected** Score	**Expected** Percentage
Peers' answers score	182	12.31%	2.4	12,06%
ChatGPT-peers' answers score	266	17.99%	3.2	16,08%
ChatGPT-teacher's score	284	19.20%	4.3	21,61%
Gemini's score	264	17.85%	2.9	14,57%
DeepSeek's score	256	17.31%	3.1	15,58%
Teachers' answers score	227	15.35%	4.0	20,10%

The highest-rated response was ChatGPT with the teacher's prompt, with 19.20% of the total score, followed closely by ChatGPT using students' prompts. With a similar score, Gemini and DeepSeek ranked after ChatGPT, while the two lowest-rated

groups were the teacher's answers and peer answers, with 15.35% and 12.31%, respectively. Summarizing the ranking scores, GenAI tools were rated significantly higher than responses obtained from peers and even from the teacher. However, the most accurate response was the one obtained using the prompt crafted by the teacher.

Regarding these results, we can confirm that students can benefit from using GenAI tools in the classroom, even if they are not the most popular ones, such as Gemini and DeepSeek. When comparing responses generated with GenAI to those without GenAI assistance by grouping all GenAI generated answers and non-Gen AI answers and calculating the percentage distribution of the scores, the results show that the Gen AI usage answers score 56.67% of the total scoring.

These results help answer **Q1** of this study. The difference was substantial, with a gap of 13.34 percentage points between the scores of Gen AI-assisted answers and those without GenAI assistance. Additionally, in Table 1 **(column 1–2)**, a 1.96 percentage point difference was observed between the highest-rated non-Gen AI response and the lowest-rated Gen AI-generated response. Although this is not a particularly large difference, it increased to around 5% when comparing ChatGPT responses with a well-structured prompt from the teacher and even reached 7% when comparing peer responses with ChatGPT-teacher responses, making a more substantial scoring gap.

In the last session, students were asked to predict the ranking before seeing the actual results. The scores they assigned differed significantly from the real results. Table 1 **(column 3–4)** shows the perceived ranking on a scale of 1 to 5 for each group.

The results highlight that students recognized ChatGPT as a powerful tool, and when used with a well-formulated prompt, they perceived it as the best option for help-seeking in problem-solving classroom sessions, with 21.61% of the total score. The second-best option, according to them, was the teacher's answer, with 20.10%, followed by the rest of the Gen AI tools. At the bottom of the ranking, they placed peer collaboration, with a 12.06% score.

This perception aligns with the actual results regarding ChatGPT with the teacher's prompt and peer responses, but it significantly differed in other aspects. During the debate in the final session, after seeing the real ranking, many students were surprised. Their reflections revealed their misconceptions, such as the lower than expected rating of the teacher's responses.

Table 1 **(columns 3–4)** contributes to answering **Q2** of the study, which focuses on students' preferences for seeking help in problem-solving classroom sessions. It highlights students' perceived importance of the teacher role while acknowledging the convenience of using a Gen AI tool such as ChatGPT as part of their learning process.

3.2 Seeking for GenAI Generated Help

When analyzing the prompts used by students when answering peer questions using ChatGPT, more than 70% of students simply copied and pasted the question, as stated in the problem proposed by the teacher, without adding any context. Fewer than 30% of students provided additional information, such as stating that they were computer science students or specifying the exact topic from the theoretical session.

When students formulated more detailed prompts, the quality of the responses improved, resulting in higher scores compared to those who simply copied and pasted

the question verbatim. Similarly, when the teacher used Gen AI, a structured prompt was crafted to generate more accurate responses, including only concise explanations and one clear example. After analyzing the different prompts, the importance of providing specific context became evident in the scores shown in Table 1 **(column 1–2)**. However, to evaluate whether students were aware of the significance of well-formulated prompts, session 4 included a debate on this topic.

During the discussion, students admitted that they usually copy and paste the problem statement directly into ChatGPT and wait for a complete response. When dealing with programming tasks, many admitted that they did not read the full explanation and instead skipped to the code solution and copied it directly.

After analyzing the results obtained from ChatGPT with a well-structured prompt and comparing them to their own obtained responses, students acknowledged the importance of adding context when prompting. Those who had already done this naturally agreed that providing context improved the efficiency of the responses.

Since the highest-scoring responses were obtained using well-structured prompts, the role of prompt formulation became a crucial part of help-seeking in problem-solving classroom sessions. Students also recognized that ChatGPT with the teacher's structured prompt was the most reliable tool for obtaining accurate responses. All this data contributes to answering **Q3** of this study, while also highlighting that the second most-rated response group was ChatGPT with students' prompts. This suggests that even without structured prompting, ChatGPT remains an effective help-seeking tool, although crafting a well-defined prompt significantly enhances response quality.

During the classroom debate in the last session about Gen AI tools, an important aspect that emerged was students' awareness of different Gen AI tools. A total of 15 different GenAI tools were introduced, most of them specializing in programming by generating code in multiple languages. However, students were generally unaware of the differences between these tools. As part of this study, the same structured prompt used was tested across different GenAIt tools, and the responses varied notably. While ChatGPT provided accurate answers, adhering to the conditions and context of the prompt, Gemini and DeepSeek did not fully consider the instructions. Instead, they generated overly detailed responses with multiple examples, failing to recognize that the responses were meant for students who did not require a fully developed solution.

When asked whether GenAI tools should be integrated into problem-solving classroom sessions, 61.11% of students responded positively, believing AI should be clearly integrated. Another 33.34% agreed but with the condition that AI should be used only in specific help-seeking scenarios such as guided problem-solving sessions. Finally, only 5.55% of students opposed arguing that the use of AI interferes with their learning process.

4 Conclusion

The results of this paper show that for certain learning scenarios, particularly in solving closed-ended tasks like software programming, GenAI can offer satisfactory help-seeking support. However, the quality of support is contingent upon the tool used and

its utilization (i.e., prompts). The dependence on tools and student-generated prompts, along with students' preferences and expectations, highlights the necessity for hybrid approaches that combine human and AI efforts in orchestrating problem-solving in the classroom. Future research should examine additional scenarios, including open-ended tasks, and investigate the deployment of orchestration systems that integrate human-AI collaboration strategies.

Acknowledgments. The work by Pompeu Fabra University authors was supported by the AEI agency of the Spanish Ministry MICIU/AEI/https://doi.org/10.13039/501100011033 (grants PID2023-146692OB-C33, CEX2021–001195-M), SGR 00930, Erasmus+ 2023–1-ES01-KA220-SCH-000157262 and KA220-HED-D8B72E6A. DHL (Serra Húnter) also acknowledges the support of ICREA Academia. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gourash, N.: Help-seeking: a review of the literature. Am. J. Community Psychol. **6**(5), 413 (1978)
2. Aleven, V., Stahl, E., Schworm, S., Fischer, F., Wallace, R.: Help seeking and help design in interactive learning environments. Rev. Educ. Res. **73**(3), 277–320 (2003)
3. Amarasinghe, I., et al.: Teacher orchestration load: What is it and how can we lower the burden? In: ISLS Community Reports, International Society of the Learning Sciences (2022)
4. Dillenbourg, P., Jermann, P.: Technology for classroom orchestration. In: New Science of Learning: Cognition, Computers and Collaboration in Education, pp. 525–552 (2010)
5. Chen, L., Chen, P., Lin, Z.: Artificial intelligence in education: a review. IEEE Access **8**, 75264–75278 (2020)
6. Giannakos, M., et al.: The promise and challenges of generative AI in education. Behav. Inf. Technol., 1–27
7. Feuerriegel, S., Hartmann, J., Janiesch, C., Zschech, P.: Generative AI. Bus. Inf. Syst. Eng. **66**(1), 111–126 (2024)
8. Bandi, A., Adapa, P.V.S.R., Kuchi, Y.E.V.P.K.: The power of generative ai: a review of requirements, models, input–output formats, evaluation metrics, and challenges. Future Internet **15**(8), 260 (2023)
9. Yan, L., Martinez-Maldonado, R., Gasevic, D.: Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 101–111 (2024)
10. Chiu, T.K., Xia, Q., Zhou, X., Chai, C.S., Cheng, M.: Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. Comput. Educ.: Artif. Intell. **4**, 100118 (2023)
11. Akgun, S., Greenhow, C.: Artificial intelligence in education: addressing ethical challenges in K-12 settings. AI Ethics **2**(3), 431–440 (2022)
12. Cukurova, M.: The interplay of learning, analytics and artificial intelligence in education: a vision for hybrid intelligence. Br. J. Educ. Technol. (2024). https://doi.org/10.1111/bjet.13514
13. Klimova, B., Pikhart, M., Kacetl, J.: Ethical issues of the use of AI-driven mobile apps for education. Front. Public Health **10**, 1118116 (2023)

14. Järvelä, S., Nguyen, A., Hadwin, A.: Human and artificial intelligence collaboration for socially shared regulation in learning. Br. J. Edu. Technol. **54**(5), 1057–1076 (2023)
15. Florido, H., Hernandez-Leo, D.: Appendices to the paper "help-seeking in problem-solving: comparing generative AI, Peers, and Teacher" (2025). https://drive.google.com/file/d/1hHOkFvTfmuxHo9AzdmrEGD70ZDqVIAKf/view

How Teacher-Provided Need Support Affects Learning Performance in Human-Computer Collaborative Teaching Environments

Shuang Li[1(✉)], Jingxi Chen[1], Jingqi Sun[2], Jing Li[1], and Xinpei Yu[1]

[1] Faculty of Education, Beijing Normal University, Beijing 100875, China
lilybnu@bnu.edu.cn
[2] Wuhan No.6 High School, Wuhan 430000, Hubei, China

Abstract. In K-12 intelligent learning environments, how teachers provide adaptive support affects student performance. This study employs a chain mediation model and learning analytics to explore how perceived need support affects learning outcomes in human-computer collaborative teaching. Results show that need support is significantly related to exercise engagement and that structure and involvement impact achievement through self-regulated learning and exercise engagement. Behavioral sequence analysis further reveals that structure and involvement shape students' behaviors and strategy during exercise on platforms. This study offers theoretical insights into teacher-computer collaboration and practical guidance for optimizing instructional design and learning support.

Keywords: Need support · Self-regulated learning · Online learning behavior · K-12 · Human-computer collaborative teaching

1 Introduction

Artificial intelligence, with functions like intelligent content analysis and assessment [8], has revitalized human-machine collaborative teaching, where teacher wisdom and machine intelligence work in synergy. This collaboration enhances classroom engagement, optimizes teaching structures and provides more efficient, precise and enriched learning support for students [10, 29]. However, research on intelligent learning systems remains focused on algorithm and model development, with limited attention to the teacher's role [4, 13]. In fact, teachers play an irreplaceable role in this environment, as they rely on experience and intuition to flexibly adjust instruction, capture non-verbal cues, provide emotional support, and foster creative learning [14]. This role is particularly crucial for K-12 students, whose cognitive and emotional development is still maturing, as it significantly enhances their learning experience and outcomes [16, 20].

Existing studies have shown that teacher-provided need support significantly influences students' intrinsic motivation [28], learning engagement [9], self-regulated learning (SRL) [1], and academic performance [7]. Based on self-determination theory, need support comprises autonomy support, structure, and involvement, all of which contribute

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 19–26, 2025.
https://doi.org/10.1007/978-3-031-99267-4_3

to enhanced learning engagement and academic achievement by fulfilling students' basic psychological needs [6, 23]. However, research on the mechanisms of teacher-provided need support in human-machine collaborative teaching environments remains limited. Most existing studies focus on theoretical strategy design, lacking empirical research from students' perspectives, and have yet to reveal how perceived need support influences students' learning behavior patterns. Furthermore, SRL is a key ability for effective learning in such environments [35], yet the mechanisms by which need support facilitates SRL remain unclear.

To address this gap, this study investigates middle school blended learning in human-machine collaborative teaching. It uses a chain mediation model to examine how need support influences SRL, online learning engagement, and learning achievement, and applies lag sequential analysis to explore behavior patterns under varying support levels. The findings aim to inform theory and guide instructional design in such environments.

2 Literature Review

Need Support is a key concept derived from self-determination theory, emphasizing the support and assistance teachers provide to students during instruction [6, 23]. It is commonly divided into three dimensions: autonomy support, structure, and involvement. Autonomy support refers to teachers understanding students' needs, interests, and emotions while providing appropriate choices. Structure involves setting clear goals, rules, and feedback. Involvement pertains to teachers expressing care, offering emotional support, and fostering positive teacher-student relationships [21, 23].

Existing research indicates that need support is closely related to academic achievement, SRL, and online learning engagement. Need support enhances learning performance not only by fostering students' SRL skills such as goal setting and strategy adjustment [1, 18, 31] but also by increasing learning duration and frequency [2, 17]. Furthermore, SRL significantly influences online learning engagement, as students with strong self-regulation can better plan and monitor their learning process, ultimately improving learning outcomes [2, 34]. This suggests that need support may indirectly affect academic achievement through SRL and online learning engagement. Additionally, different levels of need support significantly influence students' online behavioral patterns. High levels of support result in greater resource utilization and interaction frequency [15], whereas low levels lead to procrastination and social avoidance [32].

In summary, while existing research recognizes the importance of need support in SRL and online learning engagement, the independent and sequential mediating effects of these factors between need support and learning achievement in human-computer collaborative teaching environments remain unclear. Moreover, the impact of different need support levels on student behavioral patterns requires further investigation. Therefore, this study raises the following questions and hypotheses:

RQ1: In a human-computer collaborative teaching environment, does students' perceived need support influence their application of SRL strategies and online learning engagement, thereby affecting learning achievement?

H1: SRL and online learning engagement sequentially mediate the relationship between need support (autonomy support, structure, and involvement) and learning achievement.

RQ2: In a human-computer collaborative teaching environment, do different levels of perceived need support lead to variations in students' online learning behavioral patterns? What specific differences exist in these patterns?

H2: Students perceiving different levels of need support exhibit distinct online learning behavioral patterns.

3 Method

This study focuses on an English blended learning classroom in a middle school in China, combining face-to-face instruction with self-directed learning on an intelligent platform (As shown in Fig. 1). In class, teachers guide activities like listening, discussions, and presentations. During self-directed learning, students use the platform integrated with AI technology, which primarily provides three core learning functions:

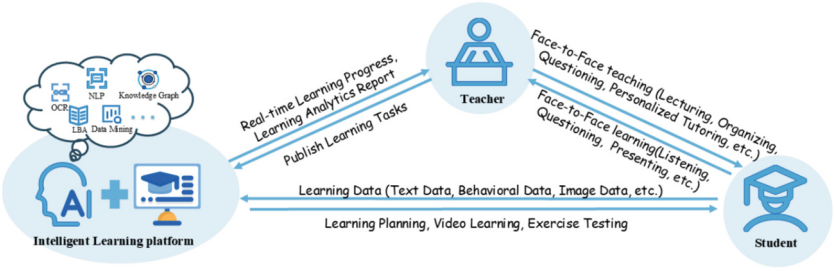


Fig. 1. Human-Computer Collaborative Teaching Model

- Learning Planning: Includes a task list (showing completed and pending tasks), task descriptions (with learning objectives, key points, and strategies), and task reports (providing AI-driven learning analytics, learning level diagnostics, and more).
- Video Learning: Allows students to control playback speed and progress.
- Exercise Testing: Supports self-assessment with AI-driven automated scoring, exercise solutions, error attribution, which provides predefined reasons for mistakes and allows students to select and record their error causes.

This study collected students' online learning behavior data, final exam scores, and self-reported questionnaire data. The valid dataset (100 seventh-graders, 63 eighth-graders; 83 males, 80 females) excluded incomplete or missing data. Informed consent was obtained from all participants and their guardians, and all personal data were anonymized for research use only. The self-reported questionnaire assessed students' perceived need support and SRL. The Need Support Questionnaire (adapted from TASC-Q [24]) and the SRL Questionnaire (adapted from OSLQ [3]) both demonstrated high reliability ($\alpha = 0.932$ and 0.897, respectively) and good structural and convergent validity (CFI > 0.95, RMSEA < 0.07). These results confirm the reliability and validity of the questionnaires used in this study.

The online learning behavior data included 16 engagement indicators across five dimensions: learning duration (course, video, and exercise), learning frequency (course, video, and exercise), focused duration (video and exercise), video adjustment frequency (rewatching, pausing, skipping forward/backward, speed-up), and exercise adjustment frequency (retaking, correcting mistakes, error attribution). Focused duration refers to the longest continuous time spent on one activity in a class session. The study also extracted students' learning behavior data from platform logs, covering interactions in the learning planning, video, and exercise sections.

4 Result

4.1 Chain Mediation Model

The correlation analysis showed that among online learning engagement indicators, only duration of exercise, frequency of exercise, focused duration of exercise, and frequency of retaking the exercise were significantly related to need support, SRL skills, and learning performance. These four indicators were selected for dimensionality reduction to represent exercise engagement.

Using Hayes' PROCESS Model 6, this study examined the chain mediation effects among need support, SRL skills, exercise engagement, and learning achievement while controlling for gender and grade. Multicollinearity diagnostics indicated that all variables included in the mediation model had variance inflation factors below 5, suggesting no serious multicollinearity issues. Figure 2 presents the path coefficients for models with structure and involvement as independent variables. In both models, the overall regression was significant ($R^2 = 0.14$, $F = 4.84$, $p < 0.001$). Bootstrap analysis showed that under structure, the indirect effects of SRL and exercise engagement were 0.06 (95% CI = [0.01, 0.12]) and 0.10 (95% CI = [0.03, 0.17]), with a combined effect of 0.01 (95% CI = [0.01, 0.07]), and a total effect of 0.16 (95% CI = [0.05, 0.07]). Under involvement, the indirect effects of SRL and exercise engagement were 0.06 (95% CI = [0.01, 0.15]) and 0.15 (95% CI = [0.05, 0.25]), with a combined effect of 0.01 (95% CI = [0.00, 0.03]), and a total effect of 0.22 (95% CI = [0.11, 0.34]). These results confirm the chain mediation effects of SRL and exercise engagement in the positive impact of structure and involvement on learning achievement. Partial support for H1 is established.



Fig. 2. Chain mediation model (Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$)

4.2 Online Learning Behavior Sequence Analysis

The study further explored behavioral sequence patterns in the exercise section based on students' perceived levels of structure and involvement, which were classified into

high and low groups using mean splits. Sequence analysis results, corrected using the Benjamini-Hochberg FDR method, show significant behavioral differences between groups, supporting Hypothesis 2 (H2).

Figure 3 shows that students with high and low structure exhibited distinct sequence patterns. The low-level group had two additional significant sequences compared to the high-level group: E2 → L3 and E3 → E5. This suggests that after submitting a test, they often checked the task report to understand their performance, or used the correct answers provided by the system to attribute and adjust errors. In contrast, the high-level group showed significant sequences E2 → E5, E5 → L3, and E3 → L1, indicating stronger self-regulation abilities. These students independently attributed errors, summarized results, and reviewed the task report after reflecting on their performance. Additionally, after reviewing the exercise analysis, they directly checked the task list to see if any tests were pending, demonstrating higher autonomy and learning continuity.

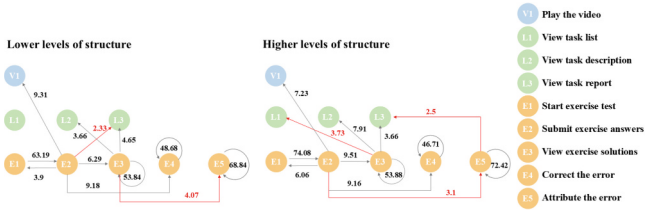


Fig. 3. Behavior sequences of students with different levels of structure

Figure 4 shows that students with high and low involvement exhibited distinct sequence patterns. The low-level group had two significant sequences, E2 → L1 and E2 → L2, indicating a tendency to check the task list and details after submitting the test, focusing on task completion rather than self-reflection or knowledge construction. In contrast, the high-level group displayed three additional significant sequences, E3 → L1, E2 → E5, and E3 → E5, reflecting stronger error attribution and reflection. After completing the test, they analyzed mistakes, reviewed tasks, and began new study rounds, demonstrating a greater emphasis on knowledge accumulation and transfer. When unable to analyze problems independently, they reattributed mistakes after reviewing exercises, summarizing problem-solving experiences to reinforce understanding, which highlights their stronger self-regulation and ability to adjust learning strategies.

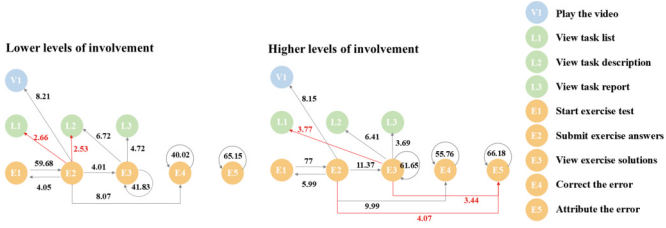


Fig. 4. Behavior sequences of students with different levels of involvement

5 Discussion and Conclusion

This study integrates a chain mediation model and learning analytics, surpassing traditional single-mediator analysis and addressing the limitations of previous studies that relied on a single research paradigm. It reveals the pathways through which structure and involvement provided by teachers in human-machine collaborative teaching influence students' learning performance, offering insights into how teachers and machines can effectively collaborate to enhance student learning outcomes.

The study identifies the correlation between exercise engagement, need support, SRL, and learning performance, emphasizing the importance of learning tasks that stimulate cognitive engagement on intelligent learning platforms for enhancing learning outcomes [33]. Additionally, the validation results of the chain mediation model show that for middle school students whose cognitive abilities and SRL skills are not fully developed, teacher cognitive guidance and emotional support effectively promote their learning [1, 11]. Cognitive scaffolding from teachers helps students bridge cognitive gaps and achieve learning goals [25], while emotional support significantly boosts students' self-efficacy, stimulates learning motivation, and fosters active participation [19]. Further behavioral sequence analysis reveals that need support significantly affects students' behavior and strategy use during exercise. Students who perceive higher levels of structure and involvement demonstrate stronger self-regulation and autonomy [19, 22], while those with lower perceived structure rely more on tools, and those with lower perceived involvement engage in more superficial learning [5]. These patterns highlight participation differences in human-computer collaborative learning and the challenges students face in translating platform feedback into improvements [12], underscoring the need for targeted support. Teachers should avoid fostering excessive reliance on intelligent tools for students with low structure and instead offer metacognitive support, such as resource learning paths, functional and cognitive feedback cues [11, 27]. For students with low involvement, enhancing interactivity and feedback is crucial to improving their sense of support and expectations, such as expressing encouragement and care regularly [26]. Additionally, current platforms' direct answer provision reinforces student dependency rather than promoting deep learning. To address this, integrating advanced generative AI, alongside teacher guidance, can foster effective interaction through step-by-step prompts and Socratic questioning [30], encouraging deeper, autonomous learning.

However, this study has several limitations. First, the sample was relatively homogeneous, which may limit generalizability. Second, the measurement of need support relied on students' self-reports, lacking validation from objective indicators of teacher behaviors. Third, individual differences such as psychological traits and prior learning experiences were not considered. Future research should include more diverse samples, combine subjective and objective indicators, and adopt experimental or longitudinal designs to clarify causal relationships among key variables.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 61977011) and the Large-Scale Longitudinal and Cross-Sectional Study of Student Development (Grant No. 2021YFC3340800).

References

1. Alten, D.C.D.V., Phielix, C., Janssen, J., Kester, L.: Self-regulated learning support in flipped learning videos enhances learning outcomes. Comput. Educ. **158**, 104000 (2020)
2. Barba, P., et al.: The importance and meaning of session behavior in a massive open online course. Comput. Educ. **146**, 103772 (2019)
3. Barnard, L., Lan, W.Y., To, Y.M., Lam, K.K.: Measuring self-regulation in online and blended learning environments. Internet High. Educ. **12**(1), 1–6 (2009)
4. Chiu, T.K.F., Moorhouse, B.L., Chai, C.S., Ismailov, M.: Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. Interact. Learn. Environ. **32**(7), 3240–3256 (2023)
5. Cents-Boonstra, M., Lichtwarck-Aschoff, A., Denessen, E., Aelterman, N., Haerens, L.: Fostering student engagement with motivating teaching: an observation study of teacher and student behaviours. Res. Pap. Educ. **36**(6), 754–779 (2020)
6. Deci, E.L., Ryan, R.M.: The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychol. Inq. **11**(4), 227–268 (2000)
7. Fujii, A.: Exploring autonomy support and learning preference in higher education: introducing a flexible and personalized learning environment with technology. Discov. Educ. **3**(1) (2024)
8. Gligorea, I., Cioca, M., Oancea, R., Gorski, A.-T., Gorski, H., Tudorache, P.:Adaptive learning using artificial intelligence in e-learning: a literature review. Educ. Sci. **13**(12), 1216 (2023)
9. Huang, L., Wang, D.: Teacher support, academic self-efficacy, student engagement, and academic achievement in emergency online learning. Behav. Sci. **13**(9), 704 (2023)
10. Jamaan, E.Z.: Arnellis: using e-learning with information theory adaptive model to increase self-regulated learning in students. J. Phys: Conf. Ser. **1940**(1), 012107 (2021)
11. Järvelä, S., Nguyen, A., Molenaar, I.: Advancing SRL research with artificial intelligence. Comput. Hum. Behav. **147**, 107847 (2023)
12. Järvelä, S., Zhao, G., Nguyen, A., Chen, H.: Hybrid intelligence: human–AI coevolution and learning. Br. J. Educ. Technol. (2025)
13. Ji, H., Han, I., Ko, Y.: A systematic review of conversational AI in language education: focusing on the collaboration with human teachers. J. Res. Technol. Educ. **55**(1), 48–63 (2022)
14. Kolchenko, V.: Can modern AI replace teachers? Not so fast! Artificial intelligence and adaptive learning: personalized education in the AI age. HAPS Educ. **22**(3), 249–252 (2018)
15. Liu, B., Luo, X., Lu, S.: A research on online teaching behavior of Chinese local university teachers based on cluster analysis. Int. J. Distance Educ. Technol. **21**, 1–20 (2023)
16. Lu, G., et al.: Analysis of perception and engagement in hybrid learning environments based on multi-agent interaction. Res. Educ. Electron. **44**(11), 82–89 (2023)
17. Luan, L., Hong, J.C., Cao, M., Dong, Y., Hou, X.: Exploring the role of online EFL learners' perceived social support in their learning engagement: a structural equation model. Interact. Learn. Environ. **31**(3), 1703–1714 (2023)
18. Ong, S.G.T., Quek, G.C.L.: Enhancing teacher–student interactions and student online engagement in an online learning environment. Learn. Environ. Res. **26**(3) (2023)
19. Prananto, K., et al.: Perceived teacher support and student engagement among higher education students – a systematic literature review. BMC Psychol. **13**, 112 (2025)
20. Rasheed, R.A., Kamsin, A., Abdullah, N.A.: Challenges in the online component of blended learning: a systematic review. Comput. Educ. **144**, 1–17 (2020)
21. Reeve, J., Deci, E.L., Ryan, R.M.: Self-determination theory: A dialectical framework for understanding socio-cultural influences on student motivation. Big Theor. Revisited **4**, 31–60 (2004)

22. Rusman, E., Nadolski, R.: Perfectly skilled: Underpinnings of an online formative assessment method for (inter)active and practice-based complex skills training in higher education (HE). Int. J. Mob. Blended Learn. **15**(2), 114–127 (2023)
23. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. Contemp. Educ. Psychol. **61**, 1–11 (2020)
24. Skinner, E.A., Belmont, M.J.: Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. J. Educ. Psychol. **85**(4), 571–581 (1993)
25. Stavredes, T.: Effective online teaching: foundations and strategies for student success. Jossey-Bass, An Imprint of Wiley, pp.288 (2011)
26. Stroet, K., Opdenakker, M.C., Minnaert. A.: Need supportive teaching in practice: a narrative analysis in schools with contrasting educational approaches. Soc. Psychol. Educ. **18**, 585–613 (2015)
27. Van Laer, S., Elen, J.: The effect of cues for calibration on learners' self-regulated learning through changes in learners' learning behaviour and outcomes. Comput. Educ. **135**, 30–48 (2019)
28. Wang, L.: Student intrinsic motivation for online creative idea generation: mediating effects of student online learning engagement and moderating effects of teacher emotional support. Front. Psychol. **13**, 954216 (2022)
29. Wang, Y., Zhu, T., Yang, S., Zheng, Y.: Human-machine collaborative teaching: motivation, essence, and challenges. Res. Educ. Electron. **8**, 51–57 (2024)
30. Wu, J., Wang, J., Lei, S., Wu, F., Gao, X.: The impact of metacognitive scaffolding on deep learning in a GenAI-supported learning environment. Inter. Learn. Environ., 1–18 (2025)
31. Xu, L., Deng, Y., Yu, C., Mo, L.: The relationship between teacher autonomy support and college students' English performance: the mediating roles of autonomy need satisfaction and self-regulated learning. Educ. Guide **12**, 57–62 (2022)
32. Zhang, S.L., Qiao, H.Y., Tang, H.L.: A study on online learning behavior and influencing factors based on MPOC. Educ. Sci. Ed. **06**, 121–127 (2022)
33. Zhang, Y., Hao, Q., Chen, B.L., Yu, H.E., Fan, F.L., Chen, Z.: Research on college students' classroom engagement and its influencing factors in smart classroom environment. China Educ. Technol. **1**, 106–115 (2019)
34. Zheng, B., Ward, A., Stanulis, R.: Self-regulated learning in a competency-based and flipped learning environment: learning strategies across achievement levels and years. Med. Educ. Online **25**(1), 1686949 (2020)
35. Zheng, B., Lin, C.-H., Kwon, J.B.: The impact of learner-, instructor-, and course-level factors on online learning. Comput. Educ. **150**, 103851 (2020)

Deep Reinforcement Learning for Engagement-Aware Question Selection in Adaptive Assessment Systems

Okan Bulut[1(✉)], Joyce Xinle Liu[1], and Hümeyra Demir[2]

[1] University of Alberta, Edmonton, AB T6G 2G5, Canada
{bulut,xinle4}@ualberta.ca
[2] Hacettepe University, Ankara 06800, Turkey
humeyra.demir@hacettepe.edu.tr

Abstract. Unlike conventional assessments, adaptive assessment systems typically require substantially fewer questions to produce an accurate estimate of the learner's ability. However, these systems work under the premise that the learner is fully engaged in the assessment, thereby putting their best effort into responding to the questions. Violations of this assumption can disrupt the question selection process, resulting in the underestimation of learners' ability levels. To address this challenge, we propose a deep reinforcement learning approach that enhances adaptive question selection by considering both learner performance and engagement. Engagement is operationalized as a latent trait, inferred from response times. We implement a deep Q-network (DQN) strategy that selects questions based on the correctness of responses and engagement status. Using real data from a large-scale digital assessment, we optimize Q-networks to improve question selection outcomes. Also, we compare the accuracy of our approach with question selection in conventional adaptive assessments. The results show that by considering learner performance and engagement, DQN yields more accurate ability estimates, especially in shorter assessments. Also, the proposed method facilitates the learners to complete the assessment within the allotted time.

Keywords: adaptive assessment systems · reinforcement learning · learner engagement · personalized assessment

1 Introduction

Over the last two decades, adaptive assessment systems have gained prominence due to their ability to tailor the assessment process to individual learners [1,4,11]. Unlike traditional assessments that administer a fixed set of questions to all learners, adaptive assessments dynamically select questions based on a learner's responses and estimated ability level. This personalized approach not only enhances the efficiency of assessments—often requiring fewer questions to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 27–34, 2025.
https://doi.org/10.1007/978-3-031-99267-4_4

achieve accurate ability estimation—but also improves the assessment experience by ensuring that learners encounter questions suited to their skill levels. However, a critical assumption underlying adaptive assessments is that learners engage fully and put forth their best effort when responding to each question. In real-world settings, this assumption may not always hold. Some learners may engage in rapid guessing or be idle for extended periods. These behaviors can compromise the accuracy of ability estimates and disrupt the question selection process. Thus, it is essential to monitor not only the accuracy of learners' responses but also their level of engagement to ensure that the most appropriate questions are selected.

To date, researchers have developed various approaches to alleviating the negative impact of low engagement on assessment results [18,23]. In this study, we propose a deep reinforcement learning (RL) approach that enhances adaptive assessments by incorporating engagement as a latent trait. Specifically, we implement a deep Q-network (DQN) to optimize question selection based on both response correctness and engagement status inferred from learners' response times on questions. The ultimate goal of this approach is to monitor the learner's engagement level in real time and use it to select the most appropriate questions, thereby enhancing the accuracy of their ability estimates.

2 Literature Review

When learners lack sufficient motivation to answer the questions, disengaged responding may manifest through behaviors such as rapid guessing (i.e., responding to questions too quickly) or idling (i.e., taking an excessively long time to respond to questions) [8,20,24]. Response times captured at the question level provide valuable insights into their test-taking behaviors and can be used to detect these types of disengaged responses. Previous studies have examined the value of incorporating response times into adaptive assessments. Response time is often used as an indicator of either speed or engagement. For the former, researchers have used response time to maximize the efficiency of adaptive assessments and manage the overall assessment duration [6,10,17]. For the latter, various approaches have been suggested to account for disengaged responding. Wise and colleagues [21,22] proposed an effort monitoring approach that detects rapid guesses and uses a scoring mechanism that excludes these responses from ability estimation. Gorgun and Bulut [9] took a more proactive approach by operationalizing test-taking engagement as a latent trait based on response times and using this information to select more optimal questions for learners.

RL is a computational framework for modeling sequential decision-making tasks, allowing an agent to learn an optimal policy through continuous interaction with its environment [16]. Given its effectiveness in sequential decision-making tasks, RL has recently been applied to enhance adaptive assessments (e.g., [13,15]). A recent study by Wang et al. [19] introduced a DQN-based strategy for question selection that outperformed traditional methods based on item response theory (IRT). Unlike conventional methods that select the next most

informative question based on a learner's current ability estimate, DQN makes it possible to consider longer-term information gains throughout the assessment. By leveraging RL, DQN can learn from the experience of other learners to predict the impact of early question choices on ability estimation. The result is a more forward-looking question selection strategy that achieves high measurement accuracy with fewer questions [19].

Building on previous work [2,9], this study further extends the idea of using a proactive approach to deal with disengaged responses. We propose a question selection algorithm that considers the learner's ability and engagement levels to select questions that are not only informative but also engaging. We leverage the advantages of the DQN strategy proposed by Wang et al. [19] and extend it to perform engagement-aware question selection.

3 Methods

3.1 Sample and Instrument

To demonstrate our RL-based approach, we used a sample of Canadian students ($n = 22,653$) who participated in a large-scale digital assessment program for grades 8 and 9. The assessment consisted of 244 multiple-choice questions focused on different subject areas, such as math and science. To create a large question bank for the current study, we assumed that the questions from these areas measured a single (i.e., unidimensional) hypothetical construct, representing learners' overall competence. For the selected sample, the dataset consisted of responses (i.e., 1=correct, 0=incorrect) and response times for all questions.

3.2 Data Analysis

To operationalize engagement as a latent trait, we followed a model-based approach introduced in earlier research [2,7]. First, we identified 20% (the threshold for rapid guessing) and 180% (the threshold for idling) of the average response time for each question. This assumes that some learners exhibit disengaged response behaviors, such as rapid guessing with very short response times, and idling without making any effort to answer the questions. Using the thresholds identified for each question, we transformed the response times into binary engagement indicators of 1=engaged if $T_{i1} \leq RT_{ij} \leq T_{i2}$ and 0=disengaged, otherwise. Here RT_{ij} is the response time of learner j on question i, and T_{i1} and T_{i2} are the thresholds for determining disengaged responding (i.e., rapid guessing and idling) for question i, respectively. For most questions in the pool, the proportion of disengaged responses was less than 20%, suggesting that the learners mostly engaged with the questions. Next, we applied the Rasch model to the dataset, which included binary indicators of engagement, to estimate two types of engagement parameters: one for the questions and one for the learners. These parameters were estimated on a logistic scale centered around zero, where higher values indicate a greater engagement demand for the questions and a higher level of engagement for the learners. To quantify learners' performance based on the

entire question pool ($n = 244$ questions), we applied the two-parameter logistic (2PL) IRT model to binary responses (i.e., 1=correct, 0=incorrect), yielding difficulty and discrimination parameters for the questions and ability parameters for the learners.

To enable engagement-aware question selection, we used a DQN-based approach [19], framing the task as a Markov decision process (MDP). In this setup, a virtual agent selects questions based on the learner's estimated ability (state), available items (actions), and the resulting Fisher information gain (reward). The DQN algorithm approximates the optimal Q-function–representing the expected cumulative information gain–via a neural network. Unlike traditional IRT-based methods, this strategy accounts for both immediate and future rewards, leading to more globally optimal item selection. Training involves learning Q-values from past learner data using experience replay and a target network. During deployment, an ϵ-greedy policy balances exploration and exploitation, selecting the item with the highest Q-value to maximize long-term information gain.

In our study, question selection is guided by the cumulative Fisher information. Two types of Fisher information are calculated for each question: ability-based Fisher information that quantifies how informative a question is regarding the learner's ability based on the correctness of responses and engagement-based Fisher information that measures question informativeness by considering learner engagement levels. The total information for each question is computed as a weighted sum of the two Fisher information values, $I_{Total} = I_{Ability} + I_{Engagement}$. The DQN algorithm, through an optimal Q-function, evaluates all available question sequences in the pool and selects the one that maximizes the long-term expected total information (i.e., Q-value). After the learner interacts with the selected question and provides a response, their ability level ($\theta_{Ability}$; based on the accuracy of the response) and engagement level ($\theta_{Engagement}$; based on whether their response time is categorized as engaged or disengaged) are updated using a scoring engine. This cycle continues until the learner reaches the end of the assessment.

3.3 Model Training and Evaluation

To evaluate the effectiveness of our approach, a post-hoc simulation study was conducted using responses and response times from the empirical dataset. Three assessment scenarios were tested: 10 questions within 10 minutes, 15 questions within 15 minutes, and 20 questions within 20 minutes. In each scenario, the assessment began with a moderately difficult question tailored to each learner. Subsequent question selection was driven by the DQN approach, which leveraged both the accuracy of the learners' responses and their response times to identify the most informative next question. For model training, 50% of the dataset (11,326 learners) was selected through stratified random sampling to ensure balanced representation across low, medium, and high levels of both ability and engagement. The remaining 50% (11,327 learners) was reserved for model evaluation. All training and testing procedures were implemented using the Python programming language within Google Colab. The same test dataset used to

evaluate the DQN-based approach was also employed to construct a benchmark adaptive assessment using an IRT-based question selection method where response accuracy served as the sole criterion for question selection, with engagement intentionally ignored. These analyses were performed using the mirtCAT package [3] within the R programming environment [14]. Bias and root mean square error (RMSE) were used to assess the recovery of the learners' true ability levels. In addition, we computed Pearson correlation (r) between the true and estimated ability levels. Lastly, we calculated completion rates (i.e., the proportion of learners who finished the assessment within the allotted time).

4 Results

Table 1 shows the simulation results for each scenario. Compared to the IRT-based selection method, the DQN-based method showed lower RMSE and higher r values across the three assessment scenarios tested in the simulation study. Although the performance gap between the two methods decreased with longer assessments (e.g., 20 questions), the DQN-based method continued to outperform the IRT-based method in terms of the accuracy of estimated ability levels. In terms of bias, the two methods yielded very similar results. The bias values were consistently positive across all assessment scenarios for both methods, suggesting that the adaptive assessments, regardless of the question selection method, overestimated the learners' true ability levels.

Table 1. Results for the DQN-based and IRT-based question selection methods.

Scenario	Method	Bias	RMSE	Correlation
10 questions within 10 minutes	DQN	0.017	0.448	0.951
10 questions within 10 minutes	IRT	0.021	0.587	0.873
15 questions within 15 minutes	DQN	0.009	0.366	0.962
15 questions within 15 minutes	IRT	0.014	0.451	0.917
20 questions within 20 minutes	DQN	0.003	0.291	0.977
20 questions within 20 minutes	IRT	0.005	0.388	0.939

Figure 1 shows the completion rates for the adaptive assessments for the learners with low ($\theta_{Engagement} < 0$) and high ($\theta_{Engagement} \geq 0$) levels of engagement. When DQN-based question selection was used, at least 93% of the learners completed the assessment, whereas the completion rate was much lower for IRT-based question selection. As the number of questions on the adaptive assessment increased, the completion rate remained the same for the DQN-based method but decreased substantially for the IRT-based method. The gap between the methods was larger for learners with lower levels of engagement.

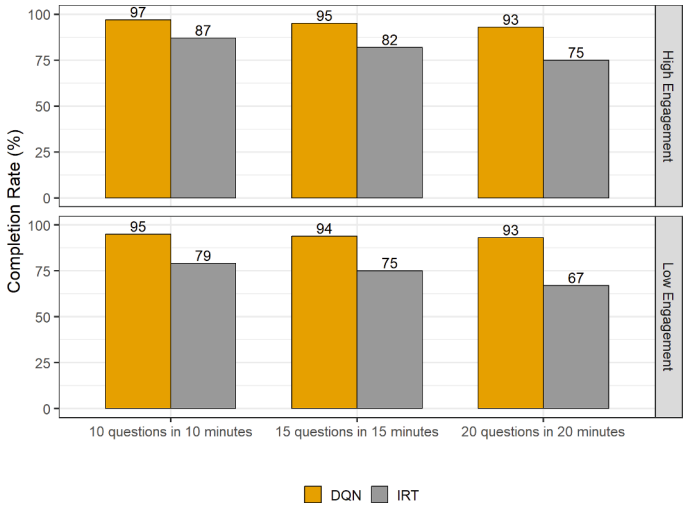


Fig. 1. Assessment completion rates.

5 Discussion

Previous research has demonstrated that RL can serve as an effective framework for improving a variety of educational applications, including personalized learning, adaptive learning, and game-based learning [5,12,13,19]. Despite these advancements, the application of RL to specifically enhance learner engagement during the assessment process has received comparatively limited attention in the literature [5]. To address this gap, the present study introduced an adaptive assessment approach incorporating an RL-based strategy to facilitate engagement-aware question selection. Our findings showed that the DQN-based question selection method could yield a lower level of error when compared with the traditional IRT-based method, particularly in shorter assessments. Furthermore, by dynamically adapting to learners' response behaviors through their responses and response times, this strategy also enhanced the efficiency of the assessment by selecting questions that were more likely to be completed in the allotted time, hence facilitating assessment completion. These results are consistent with prior research [2,9] that underscores the benefits of incorporating response time as an indicator of engagement, especially for learners prone to disengaged response behaviors (e.g., lower-ability learners or those exhibiting rapid guessing).

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

References

1. Bernardi, A., et al.: On the design and development of an assessment system with adaptive capabilities. In: Di Mascio, T., et al. (eds.) Methodologies and Intelligent Systems for Technology Enhanced Learning, 8th International Conference, pp. 190–199. Springer, Cham (2019)
2. Bulut, O., Gorgun, G., Karamese, H.: Incorporating test-taking engagement into multistage adaptive testing design for large-scale assessments. J. Educ. Meas. **62**(1), 57–80 (2024). https://doi.org/10.1111/jedm.12380
3. Chalmers, R.P.: Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. Journal of Statistical Software **71**(5), 1–39 (2016https://doi.org/10.18637/jss.v071.i05
4. Ennouamani, S., Mahani, Z.: An overview of adaptive e-learning systems. In: 2017 eighth international conference on intelligent computing and information systems (ICICIS), pp. 342–347. IEEE (2017).https://doi.org/10.1109/INTELCIS.2017.8260060
5. Fahad Mon, B., Wasfi, A., Hayajneh, M., Slim, A., Abu Ali, N.: Reinforcement learning in education: a literature review. Informatics **10**(3) (2023). https://doi.org/10.3390/informatics10030074. https://www.mdpi.com/2227-9709/10/3/74
6. Fan, Z., Wang, C., Chang, H.H., Douglas, J.: Utilizing response time distributions for item selection in CAT. J. Educ. Behav. Stat. **37**(5), 655–670 (2012). https://doi.org/10.3102/1076998611422912
7. Goldhammer, F., Martens, T., Christoph, G., Lüdtke, O.: Test-taking engagement in PIAAC. OECD Education Working Papers **133** (2016). https://doi.org/10.1787/5jlzfl6fhxs2-en
8. Gorgun, G., Bulut, O.: A polytomous scoring approach to handle not-reached items in low-stakes assessments. Educ. Psychol. Measur. **81**(5), 847–871 (2021). https://doi.org/10.1177/0013164421991211
9. Gorgun, G., Bulut, O.: Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests. Large-scale Assessments Educ. **11**(1), 27 (2023). https://doi.org/10.1186/s40536-023-00177-5
10. He, Y., Qi, Y.: Using response time in multidimensional computerized adaptive testing. J. Educ. Meas. **60**(4), 697–738 (2023). https://doi.org/10.1111/jedm.12373
11. Jose, B., Kumar, M., Udayabanu, T., Nagalakshmi, M.: Assessing the effectiveness of adaptive learning systems in K-12 education. Int. J. Adv. IT Res. Dev. **1** (2024). https://ijaitrd.com/wp-content/uploads/2024/06/Assessing-the-Effectiveness-of-Adaptive-Learning-Systems-in-K-12-Education.pdf
12. Li, X., Xu, H., Zhang, J., Chang, H.H.: Deep reinforcement learning for adaptive learning systems. J. Educ. Behav. Stat. **48**(2), 220–243 (2023). https://doi.org/10.3102/10769986221129847
13. Nurakhmetov, D.: Reinforcement learning applied to adaptive classification testing. In: Veldkamp, B.P., Sluijter, C. (eds.) Theoretical and practical advances in computer-based educational measurement, pp. 325–336. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-18480-3_17
14. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2024). https://www.R-project.org/
15. Shin, J., Bulut, O.: Building an intelligent recommendation system for personalized test scheduling: a reinforcement learning approach. Behav. Res. Methods **54**, 216–232 (2022). https://doi.org/10.3758/s13428-021-01602-9

16. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
17. Veldkamp, B.P.: On the issue of item selection in computerized adaptive testing with response times. J. Educ. Meas. **53**(2), 212–228 (2016). https://doi.org/10.1111/jedm.12110
18. Wang, C., Xu, G.: A mixture hierarchical model for response times and response accuracy. Br. J. Math. Stat. Psychol. **68**(3), 456–477 (2015). https://doi.org/10.1111/bmsp.12054
19. Wang, P., Liu, H., Xu, M.: An adaptive testing item selection strategy via a deep reinforcement learning approach. Behav. Res. Methods **56**(8), 8695–8714 (2024). https://doi.org/10.3758/s13428-024-02498-x
20. Wise, S.L.: Rapid-guessing behavior: its identification, interpretation, and implications. Educ. Meas. Issues Pract. **36**(4), 52–61 (2017). https://doi.org/10.1111/emip.12165
21. Wise, S.L.: An intelligent CAT that can deal with disengaged test taking. In: Jiao, H., Lissitz, R.W. (eds.) Application of Artificial Intelligence to Assessment, chap. 9, pp. 161–174. Information Age Publishing, Charlotte (2020)
22. Wise, S.L., Kingsbury, G.G.: Modeling student test-taking motivation in the context of an adaptive achievement test. J. Educ. Meas. **53**(1), 86–105 (2016). https://doi.org/10.1111/jedm.12102
23. Wise, S.L., Kong, X.: Response time effort: a new measure of examinee motivation in computer-based tests. Appl. Measur. Educ. **18**(2), 163–183 (2005). https://doi.org/10.1207/s15324818ame1802_2
24. Yildirim-Erbasli, S.N., Bulut, O.: The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. Educ. Res. Eval. **26**(7-8), 368–386 (2020). https://doi.org/10.1080/13803611.2021.1977152

Knowledge, Uses and Perceptions of ChatGPT in Higher Education

Christine Marquis[1], Bruno Poellhuber[2], and Normand Roy[2(✉)]

[1] Cégep de Saint-Jérôme, Saint-Jérôme, Québec, Canada
cmarquis@cstj.qc.ca
[2] Université de Montréal, Montréal, Québec, Canada
{bruno.poellhuber,Normand.Roy}@umontreal.ca

Abstract. The rapid integration of GenAI into higher education has raised an ongoing discussion about academic integrity and answer reliability. While it presents many other risks, such as biases and data privacy issues, it also offers opportunities for teachers and learners. AI literacy is therefore now deemed essential for educational stakeholders. This mixed-methods study, conducted in October 2023, involved a survey on ChatGPT uses that included an AI literacy questionnaire, based on an emerging framework of AI literacy. We analyzed semi-structured interviews with teachers and students and performed lexical analysis on open-ended responses. The findings reveal that students do not feel technically knowledgeable about GenAI and offer poor definitions of it in which they themselves have little confidence. The majority also use inappropriate strategies for conversations with ChatGPT. Despite having little technical understanding, however, they feel confident in their ability to use it. Students and educators primarily use GenAI for information retrieval and writing support, with students trusting its output more than teachers do. Although GenAI has become much better for information searches, these results point to a worryingly low student AI literacy level on the technical dimension. Both groups recognize AI's risks, particularly plagiarism, cognitive skill loss and confabulations. Both students and teachers highlight the need for formal training on GenAI, as well as institutional guidelines that are not overly restrictive. This study emphasizes the urgency of comprehensive AI literacy programs and clear guidelines at the course or institutional level to promote responsible use.

Keywords: AI Literacy · Generative AI · Teachers · Students · Higher Education

1 Introduction and Background

Since the emergence of ChatGPT in 2022, we have witnessed a proliferation of GenAI systems, including Gemini, LLaMA and DeepSeek, which rely on large language models (LLMs) and natural language processing techniques to both generate coherent text and simulate real human conversation. Interest in generative artificial intelligence (GenAI) in higher education has surged, with concerns about plagiarism and academic integrity at the forefront of discussions, but many other ethical concerns about GenAI have also

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 35–42, 2025.
https://doi.org/10.1007/978-3-031-99267-4_5

emerged in public spaces: biases, hallucinations, confidentiality of personal information, misinformation [1], environmental impact, etc.

To mitigate plagiarism risks, educational institutions have responded in various ways. Some have revised their academic policies on plagiarism. Certain institutions have adopted AI-generated text detection tools, despite ongoing debates regarding their reliability and accuracy and the problems of false positives and false negatives. In spite of its challenges, GenAI is widely regarded as having the potential to transform education in positive ways [2]. It may enhance student learning by providing personalized feedback, supporting comprehension and facilitating tasks such as writing, revising and translating texts. For educators, GenAI can serve as a valuable tool for lesson planning, instructional design and curriculum development, leading to more efficient and adaptive pedagogical approaches. Some universities have introduced guidelines (rather than policies) on how to use AI productively and effectively, acknowledging the fact that GenAI is already used by both teachers and students and that it is more productive to guide that to restrict. In this context, to ensure that students, educators and policymakers have the knowledge and skills to leverage GenAI responsibly, AI literacy is deemed essential [2, 3].

One of the most cited definitions of AI literacy is *a set of competencies that enables 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* [4]. Six core AI literacy competencies have been identified [4]. Teachers must become familiar with the uses of AI in education to effectively wield these tools and understand the associated risks [5]. From our initial literature review, we retained three dimensions for AI literacy: technical knowledge, ethical issues and general ChatGPT use. A scoping review adds more dimensions to this framework: critical thinking about AI, use of AI for learning and teaching, and human agency in AI use, an affective dimension related to the need to keep human decisions in the loop [6].

The objective of this part of the research is to measure teachers' and students' AI literacy and self-efficacy in ChatGPT use and to document their uses. We also aim to describe teachers' management of GenAI based on students' desires and institutions' management of GenAI based on teachers' desires. In light of historical results showing a higher level of technological confidence in men, we will also investigate gender differences in these perceptions and uses.

2 Method

2.1 Research Design and Participants

This research is a large survey based on a mixed methodology using questionnaires and group interviews carried out in October 2023. It targeted the entire population of teachers at two colleges and one university, as well as the entire student population at both colleges and 30% of the student population at the university. In total, we collected 2,407 valid student responses (1,055 college students and 1,352 university undergraduates, master's or PhD students; 709 men,1,606 women, 39 non-binary and 53 missing gender data; average age of 23). For teachers, we collected 322 responses (102 at the college level and 220 at the university level (including 76 who are only teaching); 145 men, 163 women, 3

non-binary and 11 missing gender data; average age of 45.8). From these questionnaires, 32 students and 22 teachers were recruited for interviews. Ethics approval applications were submitted to the ethics committees of the three participating institutions, and an ethics compliance certificate was issued by each of them. Before completing the fully anonymous questionnaire, the participants provided informed consent after reading a simplified information and consent form on the first page of the questionnaire, which also included a link to the full information and consent form. The participants who wished to participate in the interviews provided their names in response to a question in the questionnaire and subsequently received, completed and signed a detailed free and voluntary consent form. This form guaranteed them confidentiality through the use of a data de-identification procedure that replaced personal identifiers with alphanumeric codes.

2.2 Data Collection and Analysis

The measurement of AI literacy was initially based on a 33-item questionnaire based on some existing questionnaires [7–10], adapted by [11], and aligned with the three dimensions identified (technical knowledge, ethical issues and general AI use). The results of exploratory factor analyses carried out with teachers and students led to a solution comprising 21 items covering three factors: AI technical knowledge (12 items – e.g., I'd be able to explain to someone what training data is), general use of AI (4 items – e.g., I use generative artificial intelligence tools to help me create textual content), and ethical issues (5 items – e.g., I'm always aware of possible abuses of AI). In both cases, a scale of personal efficacy in using ChatGPT was also developed, and the factorial solution converged on the same factor with 7 items. After checking the scales for normality, we applied one-factor ANOVAs to compare teachers with students on these scales and to compare genders (men/women). We treated "non-binary" and "prefer not to answer" as missing because the numbers were below 30. We ran lexical analysis on two open-ended questions: *Can you briefly explain what ChatGPT is and how this tool works? When you submit a question or request to ChatGPT, what are your strategies for getting satisfactory answers?* Content analysis was carried out on two open-ended questions: *Describe the three main ways you have used ChatGPT. For each use, first describe how you used the tool, then evaluate the quality of the responses it provided. In your opinion, what are the main issues, risks or challenges posed by ChatGPT for colleges and universities?* Two independent coders coded all the answers, achieving an interjudge agreement of 85% [12]. We held eight group interviews with students (3 to 5 participants per interview) and ten group interviews with teacher (two or three participants per interview) on their uses of ChatGPT, their concerns about ChatGPT and their preferences for ChatGPT management by teachers or institutions. A content analysis was carried out on the interview transcriptions, with an intercoder agreement over 85% on 25% of the interviews.

3 Results

3.1 General Knowledge About AI

At the time of the study, 43% of students reported not using any ChatGPT. In the interviews, they justified not using GenAI out of fear of being accused of cheating by professors and due to the lack of transparency of the AI algorithms and companies.

Various definitions of ChatGPT were provided by students and teachers, ranging from very simple to complex, with many being very general and others containing errors. The teachers' confidence levels in their definitions ($M = 3.47$) were significantly higher than students' confidence ($M = 3.04$; $F(1, 1675) = 37.77$, $p < .001$).

The results of the lexical analysis on the strategies used by students to obtain satisfactory answers show four classes of answers: Class 1: I don't use it (11.8%); Class 2: Clear, precise, simple questions (52.6%); Class 3: Keywords, essays (12.1%); Class 4: Details, define, explain, specify, ask (23.4%).

Table 1 shows the main results of an ANOVA concerning AI literacy, ChatGPT self-efficacy and intent to adopt ChatGPT, contrasting students and teachers. It reveals that teachers differ significantly from students on technical knowledge, ethical aspects and intent to adopt. Students use GenAI tools to create text, images or other output more than teachers do, but they have less technical knowledge about GenAI.

Table 1. AI literacy dimensions ANOVA comparing students to teachers

Dimension	n	Mean	SD	F	p
1. Technical knowledge				56.303	< .001
Teachers	190	3.224	1.44		
Students	1225	2.490	1.22		
2. General AI use				1.912	0.167
Teachers	190	2.332	1.36		
Students	1251	2.198	1.23		
3. Ethical issues				5.061	0.025
Teachers	191	4.682	1.03		
Students	1245	4.514	0.95		
ChatGPT self-efficacy				1.339	0.247
Teachers	174	3.453	0.92		
Students	1092	3.372	0.84		
Intent to adopt				7.379	0.007
Teachers	168	2.930	1.38		
Students	1044	2.630	1.33		

Table 2 shows the main results concerning AI literacy, ChatGPT self-efficacy and intent to adopt ChatGPT by gender, combining both faculty and students.

Table 2. AI literacy dimensions ANOVA comparing genders

Dimension	n	Mean	SD	F	p
1. Technical knowledge				222.30	< .001
Men	432	3.278	1.34		
Women	931	2.250	1.10		
2. General AI use				35.21	< .001
Men	443	2.510	1.41		
Women	946	2.091	1.13		
3. Ethical issues				23.44	< .001
Men	448	4.704	0.87		
Women	935	4.440	0.99		
ChatGPT self-efficacy				78.51	< .001
Men	405	3.673	0.76		
Women	818	3.227	0.86		
Intent to adopt				32.99	< .001
Men	399	2.990	1.37		
Women	774	2.520	1.30		

3.2 GenAI Uses

According to the analysis of the open-ended questions, the most common student uses of ChatGPT are, from most frequent to least frequent: information search, studying subject matter related to their course, idea generation, text revision, text writing, text summarizing, translation, general exploration and coding. The most common uses of ChatGPT by teachers are similar: information search, text revision, text writing, translation, idea generation, testing of existing assessment in their courses, general exploration, pedagogical activities planning, summarizing and coding. While many students consider the quality of answers to be very good for their two main uses, teachers are much more critical of answer quality for information search.

3.3 Concerns

The main concerns of both faculty and students emerging from both the interviews and the analysis of the open-ended questions converged closely: plagiarism, cognitive loss, unreliable answers, quality of learning, assessment of learning and intellectual property.

3.4 Desired GenAI Management Strategies

When asked about the GenAI management strategies they would like teachers to adopt, students primarily wanted teachers to teach students how to use it (n = 18), allow its use

(n = 18), adapt assessments accordingly (n = 8), integrate ChatGPT into their courses (n = 7) and train themselves in its use (n = 6). For instance, Malika, advocating for its acceptance, stated, "I was told not to use it. I said, 'Oh, but what the heck ?' … There's no point in forbidding it because people like me will only be more motivated to explore it." In teacher interviews, the most commonly reported management strategies were addressing plagiarism (n = 9), allowing its use (n = 9), training students (n = 8), adapting assessments (n = 7), supervising usage (n = 6) and limiting its use (n = 4). One teacher explained, "I allow students to use it simply because I cannot control how they use it." Regarding the GenAI management strategies they expected institutions to promote, teachers primarily wanted institutions to train educators (n = 10), regulate its use (n = 8), respect academic freedom (n = 6) and allow its use (n = 4).

4 Discussion and Conclusion

Generally, our results point to low levels of knowledge about ChatGPT and low AI literacy in students. Compared to teachers, our results show a lower level of student technical AI literacy and less confidence in their definition of ChatGPT but a similar, medium level of ChatGPT self-efficacy. Students also report inappropriate strategies for interacting with it. Second, with regard to the main student and teacher use, information research, students rated the quality of its responses as good or very good, while teachers were more critical. This raises concerns about the students' capacity to critically assess AI-generated content. While newer tools have significantly improved LLMs for information search (e.g., Gemini's or ChatGPT's deep search), this use was not appropriate in 2023, due to LLMs' tendency to hallucinate. Third, even though both students and teachers expressed a high awareness of AI's potential ethical risks (high ethical issue subscale mean and qualitative results on plagiarism, cognitive skill loss and confabulations), their limited technical AI literacy and poor understanding of AI mechanisms may prevent them from fully grasping the ethical risks of GenAI and thus using it properly and responsibly.

There are considerable gender disparities in AI literacy scales and adoption intentions. Male students exhibit higher scores on every scale. This gap may stem from long-standing gender differences in technology acceptance and use which may be attributed to differences in self-perception rather than performance. Some other studies also highlight these differences [13]. AI literacy programs should focus on better educating women about GenAI and its potential, an avenue that emerges as a promising way to reduce gender bias and encourage women to pursue AI careers [14].

When asked how GenAI should be managed, both students and teachers emphasized the need for training as a top priority. While students believe its use should be allowed without restriction, teachers argue that institutions should permit GenAI usage but provide clearer guidelines to regulate it. These perspectives highlight the importance of having clear guidelines at both the course and institutional levels to encourage responsible AI use within clear boundaries.

The strengths of this study lie in the large number of respondents and the variety of data collection tools, enabling triangulation. These findings underscore the urgent need to implement a large-scale AI literacy initiative to ensure informed and responsible uses

of ChatGPT and GenAI in higher education. Students should be the first target of this initiative, but AI literacy programs should also be aimed at teachers and professionals who train teachers and students (librarians, instructional designers), as well as leaders, so they can establish appropriate guidelines to promote useful and responsible educational uses rather than focusing mainly on plagiarism and legal compliance.

The fact that AI literacy and adoption intentions are significantly lower among women is concerning. This situation could lead to an even greater gender-based digital divide than the one that already exists in digital technologies, despite the fact that future careers in programming and AI should be seen as offering opportunities for women as well. Gaining a better understanding of women's preferred uses of AI, compared to those of men, could help design activities specifically targeted at women. With this in mind, it would be appropriate to draw inspiration from initiatives such as *Girls Who Code* to offer training activities specifically designed for women, with the aim of sparking and sustaining their interest in AI within their professional fields.

The limitations of this study are related to the fact that the data collected are primarily perceptual rather than observational. Another limitation stems from our decision to treat responses from non-binary participants as missing data, due to the small number of respondents. An analysis that includes diverse gender identities would provide a more accurate account of differences attributable to gender rather than sex and could inform practice for designing even more inclusive training initiatives.

Further work is required to document the evolution of perceptions and uses of GenAI, AI literacy and the willingness of both students and faculty to be trained on AI literacy. An initiative of this kind is currently being carried out in the institutions that participated in the study that is the basis for this paper.

Acknowledgments. This research was funded by The Fonds de recherche du Québec – Société et culture (FRQSC) Soutien à la recherche au collégial program (https://doi.org/10.69777/340068), the Ministère de l'enseignement supérieur du Québec (Partenariat UdeM program) as well as by a financial contributions of the two colleges involved (Cégep de Saint-Jérôme and Cégep régional de Lanaudière à Terrebonne).

Disclosure of Interests. The authors declare they have no competing interests.

References

1. Shoaib, M.R., Wang, Z., Ahvanooey, M.T., Zhao, J.: DeepFakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. In: 2023 International Conference on Computer and Applications (ICCA), pp. 1–7. IEEE (2023). https://doi.org/10.1109/icca59364.2023.10401723
2. UNESCO. Guidance for generative AI in education and research. UNESCO (2024). https://unesdoc.unesco.org/ark:/48223/pf0000386693
3. Yan, L., Greiff, S., Teuber, Z., Gašević, D.: Promises and challenges of generative artificial intelligence for human learning. Nat. Hum. Behav. **8**(10), 1839–1850 (2024). https://doi.org/10.1038/s41562-024-02004-5
4. 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). https://doi.org/10.1145/3313831.3376727

5. Wilton, L., Ip, S., Sharma, M., Fan, F.: Where Is the AI? AI Literacy for Educators. Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium, pp. 180–188 (2022). https://doi.org/10.1007/978-3-031-11647-6_31
6. Tremblay, C., Roy, N., Poellhuber, B., Sénécal, A.M., Cuerrier, M., Lapierre, H.: Intégration de l'IA au postsecondaire. Presentation at the Journée du numérique en éducation, Quebec (2024)
7. Chai, C.S., Wang, X., Xu, C.: An extended theory of planned behavior for the modelling of Chinese secondary school students' intention to learn artificial intelligence. Mathematics **8**(11), 2089 (2020). https://doi.org/10.3390/math8112089
8. Kim, S.-W., Lee, Y.: The artificial intelligence literacy scale for middle school students. J. Korea Soc. Comput. Inf. **27**(3), 225–238 (2022). https://doi.org/10.9708/jksci.2022.27.03.225
9. Wang, B., Rau, P.-L.P., Yuan, T.: Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. Behav. Inf. Technol (2023). https://doi.org/10.1080/0144929X.2022.2072768
10. Zhao, L., Wu, X., Luo, H.: Developing AI literacy for primary and middle school teachers in China : based on a structural equation modeling analysis. Sustainability **14**(21), 21 (2022). https://doi.org/10.3390/su142114549
11. Lepage, A.: Étude de l'adoption des principaux types d'usages de l'intelligence artificielle par les enseignants et enseignantes du postsecondaire. [Doctoral dissertation, Université de Montréal] (2024). https://papyrus.bib.umontreal.ca/xmlui/handle/1866/40548
12. Miles, M.B. Huberman, A.M.: Qualitative Data Analysis: An Expanded Sourcebook. SAGE Publications (1994)
13. Møgelvang, A., Bjelland, C., Grassini, S., Ludvigsen, K.: Gender differences in the use of generative artificial intelligence Chatbots in higher education: characteristics and consequences. Educ. Sci. **14**(12), 1363 (2024). https://doi.org/10.3390/educsci14121363
14. Shah, S.S.: Gender bias in artificial intelligence: empowering women through digital literacy. J. Artif. Intell., 1 (2024). https://doi.org/10.70389/pjai.1000088

Inclusive Data Literacy: UDL and AI-Assisted Data Storytelling for BVI Students

Marina Buzzi[1], Barbara Leporini[2], Angelica Lo Duca[1(✉)], Veronica Punzo[3], and Daniela Rotelli[4]

[1] IIT-CNR, Pisa, Italy
{marina.buzzi,angelica.loduca}@iit.cnr.it
[2] University of Pisa & ISTI-CNR, Pisa, Italy
barbara.leporini@unipi.it
[3] Scuola Superiore Sant'Anna, Pisa, Italy
veronica.punzo@santannapisa.it
[4] Sorbonne Université, Paris, France
daniela.rotelli@lip6.fr

Abstract. Blind and Visually Impaired (BVI) students face significant challenges in accessing and understanding data. Traditional educational resources often rely on visual aids, making it difficult for BVI students to develop data comprehension skills. This study proposes a new approach using Generative Artificial Intelligence, Universal Design for Learning, and Data Storytelling techniques to help BVI students understand data through narratives. By converting complex data concepts into interactive and adaptive narrative-based experiences, our aim is to reduce the accessibility gap and enhance cognitive engagement. Preliminary experiments suggest the feasibility of the proposed methodology. More testing with end users is needed to confirm its effectiveness.

Keywords: K-12 Data Storytelling · Blind and Visually Impaired Students · Generative AI · Data Literacy · Universal Design for Learning

1 Introduction

In today's digital society, data literacy - defined as the ability to articulate information needs, retrieve, evaluate, and manage data [12] is an essential skill, projected to be paramount by 2030 [10]. However, blind and visually impaired (BVI) students face unique challenges in developing these skills due to the significant reliance of traditional data education on visual representations [5]. In Italy, since the late 70 s students with disabilities have attended common classes (from primary to high school) in which curricular teachers are assisted by support teachers [9]. In the 2023/2024 school year, more than 359,000 students with disabilities were enrolled in Italian schools, including around 12,000 BVI

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 43–51, 2025.
https://doi.org/10.1007/978-3-031-99267-4_6

students [21]. Regrettably, despite progress in accessibility technologies, visually impaired students lack adequate educational resources to comprehend and engage with data.

Universal Design for Learning (UDL) addresses this need by promoting accessible learning through various representation, action, and expression methods, key elements of multimodal learning that support inclusion [30]. Storytelling has long been recognised as an effective pedagogical tool as it enhances engagement and comprehension by placing abstract concepts in a familiar and relatable context [36]. Like conventional storytelling, data storytelling (DS) employs narrative structures and visual representation to present data insights in an engaging and comprehensible manner [11]. Unfortunately, visual representation is not accessible to BVI students. One promising avenue is using GenAI to create accessible learning experiences by converting complex information into narrative-driven and multimodal experiences that align with the needs of BVI students.

Many works in the literature combine UDL and AI [17,33], UDL and storytelling [3], or AI and DS [23,24,43]. To our knowledge, no previous research has explored the integration of UDL, DS, and GenAI in a systematic methodology for BVI students. This paper explores the potential of UDL and AI-assisted DS to improve data literacy for BVI students. We propose a methodology that takes as input a general educational programme to introduce students to data literacy and leverages GenAI to convert the programme into data-driven stories for BVI students. The stories generated consider the educational needs of a specific BVI student and the available assistive technologies. Beyond generating narratives, our methodology also suggests the most effective modalities to communicate these stories, such as auditory, sonification, haptic feedback, or multisensory storytelling techniques. This adaptive approach ensures an engaging, understandable and tailored educational experience. Preliminary experiments have been carried out to verify the feasibility of the proposed approach.

2 Universal Design for Learning for BVI Students

Based on the *Design for All* philosophy of the European Institute for Design and Disability [1], Universal Design for Learning (UDL) is a framework developed by CAST [40] that prioritises student-centred design and promotes educational diversity through the effective use of digital technology [4,14]. Introduced in 2008 and continuously updated in response to feedback from practitioners and new research, the UDL guidelines emphasise the importance of designing learning goals, methods, and materials that are valid for all students. With the aim of creating an inclusive learning environment that allows all students to reach their full potential, this framework acknowledges student differences by incorporating neuroscientific and pedagogical principles to provide multiple means of representation, action, expression, and engagement. Research by [6,30] demonstrates that UDL effectively improves accessibility and increases student participation in schools. UDL is based on the International Classification of Functioning, Disability, and Health (ICF) [13]. Although ICF focusses on barriers to learning

and participation, UDL addresses curriculum-related barriers, highlighting the importance of context in achieving school inclusion and enabling teachers to effectively support BVI students by fostering cognitive potential [2,29]. Early interventions for BVI students should consider their observational skills while promoting the development of essential competencies for future learning [7] and prioritising hands-on experiences [16]. An inclusive approach emphasises communication through multiple sensory channels and adapts materials to meet individual needs . However, although in preschool BVI students develop sensory and motor skills through touch, hearing, taste, and smell, in primary school, reliance on visual and verbal communication can complicate understanding. Hence, the creation of teaching materials that are accessible and specialised to meet the individual needs of each student represents a critical challenge.

3 Data Storytelling for BVI Students

Data storytelling (DS) combines data, narrative and visualisation to communicate complex information in an accessible and engaging way [35]. For BVI students, DS offers a powerful alternative to traditional data literacy education, which frequently depends on visualisations such as charts and graphs [34] that represent barriers for BVI students, even when adapted through screen readers or tactile methods graphics [22]. In contrast, DS can leverage auditory and tactile modalities, enabling BVI students to engage with data through multisensory narratives. Research has shown that stories make data more relatable by providing context and emotional engagement, which fosters comprehension and retention [26]. For example, storytelling enables learners to connect data with personal experiences, make sense of complex patterns, and draw meaningful conclusions [27]. Explanatory visualisations, supported by narrative elements, can significantly improve accessibility by helping learners grasp key ideas without requiring advanced data interpretation skills [28]. AI potentially can enhance DS by tailoring the complexity of stories, adjusting sensory modalities, namely, providing sensory alternatives to BVI students for accessing content, and providing real-time feedback, creating context-aware narratives that optimise learning outcomes. AI-assisted tools can also automate storytelling processes, reducing the cognitive and logistical burden on educators [24].

We identify four types of data story: change over time, comparison between entities, zoom stories, and spatial change [35]. A data story relies on a clear narrative structure to guide the audience through complex information and insights. Several models from fiction and cinema have been adapted for DS, including the three-act structure, which organises the narrative into a setup, confrontation, and resolution [25]. Freytag's Pyramid extends this model by emphasising rising action, climax, and falling action to create narrative tension and highlight key insights [44]. Similarly, the Data Storytelling Arc simplifies Freytag's Pyramid by focussing on setting, rising insights, an "Aha!" moment, and actionable outcomes [8]. The Hero's Journey presents the data narrative as a transformative quest structured around the departure, initiation, and return phases [42].

4 The Proposed Methodology

The proposed early stage methodology is based on the expertise of one of the authors, who teaches Educational Technology for Inclusion at the university level within a Primary Teacher Education programme. This background ensures that the approach is informed by both pedagogical theory and practical experience in training future primary school teachers to address diverse learning needs.

Starting from these premises, in this work, we aim to tackle the issue of data literacy programmes for primary school BVI students in Italy by proposing a methodology that, combining UDL and DS, enables teachers to customise the general educational programme for the entire classroom to meet the unique needs of the BVI students, thereby creating an accessible and personalised learning path. As illustrated in Fig. 1, this transformation is achieved through AI-driven automation.

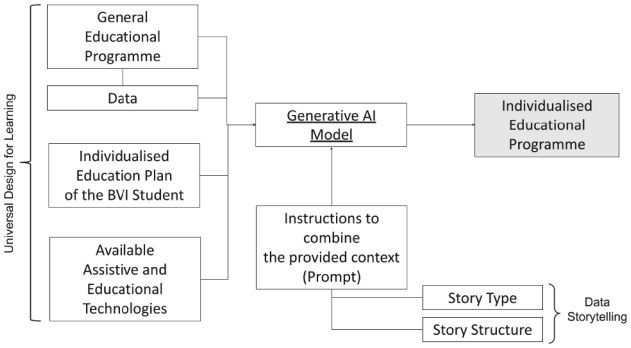


Fig. 1. The proposed methodology.

The UDL's emphasis on flexibility and inclusion is reflected in the integration of three key components: *i)* the General Education Programme (GEP) with the specific data used as examples; *ii)* the Individualised Education Plan (IEP) of the BVI student; *iii)* the Available Assistive and Educational Technologies (AAET). GenAI intervenes to combine UDL and DS principles, hence personalising the educational programme to meet each student's individual needs.

GEP. According to the International Standard Classification of Education [20], educational programmes are intended to improve the general knowledge, skills, and competencies of learners to establish a foundation for lifelong learning. Primary education programmes are generally structured to provide students with key reading, writing, and mathematics abilities while creating a solid foundation to learn and understand core knowledge areas [41]. The GEP includes data, which refers to information provided as input to the system for the purpose of generating a narrative. For example, students' preferences or event details, such as the number of animals, colour, and type of food.

IEP. Student's IEP, created for each student with special educational needs, examines key environmental and personal factors such as relationships, socialisation, communication and autonomy to promote inclusive intervention [18]. Taking into account various classroom dynamics, it enables effective pedagogical planning that respects different ways of expression and interaction.
AAET. Within the framework of educational technologies, understood as digital tools and resources designed to improve and transform teaching and learning processes [15], assistive technologies are specifically designed to support individuals with disabilities or special needs to promote their health, well-being, inclusion, and participation [39]. Braille displays and speech synthesis software allow BVI students to access learning materials [32].

GenAI integrates UDL and DS: by analysing UDL content, including the GEP, the BVI student's IEP, and the AAETs, GenAI creates a personalised educational programme tailored to the needs of BVI students. UDL personalises content and methods for individual learning requirements, while DS transforms material into engaging narrative-based lessons that enhance data literacy. In addition, GenAI recommends suitable assistive technologies to improve accessibility and personalised learning experiences.

5 Preliminary Experiments

To test the feasibility of the proposed methodology, we performed two preliminary experiments using OpenAI ChatGPT Plus in versions 4.0 and 4.5 as GenAI models. We supplied each model with various context parameters (Sect. 4) using the specific functionality offered by ChatGPT for document uploads. Initially, we also attempted to use Microsoft Copilot Pro; however, its limitation on simultaneous document uploads led us to concentrate solely on ChatGPT. We conducted a case study on a first-grade class in a primary school in Italy. All context documents are available on GitHub [38]. The GEP includes three progressive lessons: (1) observation and data collection, (2) organisation and visual representation of the data, and (3) interpretation of the data and its use to build a narrative. The BVI student's profile was derived from the IEP, appropriately anonymized, which establishes personalised learning objectives and competencies. The IEP concerns a 6-year-old completely blind girl named Cecilia (nickname for privacy issues). As assistive and didactic technologies, we included a screen reader, a rubber board (tactile writing board), clay, and an accessible tablet. As data, we used data related to animals (as specified in the GEP), and the goal was to make children able to classify animals based on colour, size, and diet. Starting from this context, students had to build stories of the comparison between entity types following the three-act structure (Sect. 3). We used the following prompt to instruct ChatGPT to generate the adapted programme:

Adapt the attached programme, titled Data Literacy Programme, for Grade 1 to the attached profile of the Blind and Visually Impaired student, titled IEP Cecilia, using the assistive and didactic technologies specified in the attached file, titled Available Technologies. In adapting the programme, generate a data-driven

story that uses the data provided in CSV format, is based on entity comparison, and follows a three-act structure.

The generated programmes, whose results are discussed in the following section, are available on GitHub [38].

6 Discussion and Future Work

The generations of stories through ChatGPT differed according to the model used. The 4.0 model generated four lessons, where only the fourth was a data story; the 4.5 model instead generated three lessons already organised as a data story following the three-act structure. The two versions of the programme adapted for Cecilia have similar approaches but significant differences in the tools used. The GPT4.5 version focusses on a three-act structured narrative and uses mainly digital technologies such as the accessible tablet with screen reader and voice dictation function, the tactile tablet for raised line graphs and clay models for animal classification. In contrast, the GPT4.0 version takes a more sensory and spatial approach, employing stackable blocks to build tactile bar graphs, symbols, and textures to represent data, haptic feedback on the tablet, and voice recording as the final form of expression. While the former focusses on the shared construction of a data-driven story, the latter emphasises autonomous, physical space exploration, and personal recorded narration. Thus, GPT 4.5 seems to be more suitable and effective. From a practical and educational perspective, the proposed methodology contributes to inclusive classroom practices by providing teachers with customised tools and adaptive strategies. In particular, combining UDL with AI-assisted data storytelling can allow educators to tailor the learning experiences of BVI students without isolating them from their peers. This approach also improves early data literacy by encouraging essential skills such as critical thinking and problem solving, thereby promoting long-term academic growth of BVI students and increasing their independence in daily life. However, the study presents some limitations that require consideration in future research. The preliminary experiments should be validated in real-world classroom scenarios that include UDL factors of increasing complexity. To ensure equitable and secure learning environments, ethical implications, such as the inherent biases of GenAI models and privacy concerns about personal data, must be further investigated [31,37]. To mitigate risks related to biassed or overly simplified content generation, teachers will review all AI-produced stories before use. The system is designed as a co-creative assistant rather than an autonomous generator, ensuring human oversight in every step. Field testing is planned for summer 2025 in collaboration with the Italian Association of Blind and Visually Impaired [19]. During these summer camps, we aim to evaluate the effectiveness of DS in improving data literacy with about 12–16 K-12 BVI students. To this end, we will define some evaluation metrics and conduct a comparative study involving two groups of students: the first group will follow a

traditional educational path without using DS. In contrast, the second group will engage with the AI-assisted DS approach. Both groups will complete pre- and post-programme questionnaires to assess learning outcomes and measure any improvements in data literacy. In this paper, we used a generic GenAI model. We plan to apply retrieval-augmented generation techniques and multi-agent architectures to improve performance. This preliminary study suggests the feasibility of combining UDL, DS, and GenAI for inclusive education. If confirmed, the proposed approach could become a replicable methodology in other contexts.

Acknowledgements. This study is part of the PRIN project 2022HXLH47 "STEMMA -Science, Technology, Engineering, Mathematics, Motivation and Accessibility" (funded by the European Union - Next Generation EU, Mission 4 Component C2 CUP B53D23019500006).

References

1. The EIDD Stockholm Declaration (2004). https://tinyurl.com/2jtt3hsc
2. Ainscow, M.: Promoting inclusion and equity in education: lessons from international experiences. Nord. J. Stud. Educ. Policy **6**(1), 7–16 (2020)
3. Alharbi, S., Newbury, P.: Improving student engagement and satisfaction using universal design for learning and storytelling. In: European Conference on E-Learning. pp. 559–XIII. Academic Conferences International Limited (2021)
4. Aral, S.: The hype machine: How social media disrupts our elections, our economy, and our health–and how we must adapt. Crown Currency (2021)
5. Bhat, S.A., Alyahya, S.: Infographics in educational settings: a literature review. IEEE Access **12**, 1633–1649 (2023)
6. Cook, S.C., Rao, K.: Systematically applying udl to effective practices for students with learning disabilities. Learn. Disabil. Q. **41**(3), 179–191 (2018)
7. Diasse, M.D., Kawai, N.: Evidence-based practices for students with visual impairments in regular settings. Int. J. Spec. Educ. **39**(2), 75–86 (2024)
8. Dykes, B.: Effective data storytelling: how to drive change with data, narrative and visuals. John Wiley & Sons (2019)
9. Teacher education for inclusive education. https://tinyurl.com/3wfeh3j5
10. Data Literacy: The Upskilling Evolution. https://tinyurl.com/yfna2729
11. Few, S.: Show me the numbers: Designing tables and graphs to enlighten/
12. DigComp Framework. https://tinyurl.com/2s3e6p2b
13. International Classification of Functioning, D., Health (ICF). https://tinyurl.com/45b7wvvb
14. Gardou, C.: La société inclusive, parlons-en! Érès (2012)
15. Granić, A.: Educational technology adoption: a systematic review. Educ. Inf. Technol. **27**(7), 9725–9744 (2022)
16. Hunaepi, H., Suma, I.K., Subagia, I.W.: Curiosity in science learning: a systematic literature review. J. Essent. Competencies Educ. **3**(1), 77–105 (2024)
17. Hyatt, S.E., Owenz, M.B.: Using universal design for learning and artificial intelligence to support students with disabilities. College Teaching, pp. 1–8 (2024)

18. Il Decreto interministeriale. https://tinyurl.com/b2s2zesc
19. Unione Italiana dei Ciechi e degli Ipovedenti. https://www.uiciechi.it/
20. International Standard Classification of Education (ISCED). https://tinyurl.com/3w5hydc6
21. ISTAT. https://shorturl.at/WAeAH
22. Kakehi, B., Iwamura, M., Minatani, K., Kise, K.: Grouping effect for bar graph summarization for people with visual impairments. In: Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (2024)
23. Li, H., Wang, Y., Liao, Q.V., Qu, H.: Why is AI not a panacea for data workers? an interview study on human-AI collaboration in data storytelling. IEEE Trans. Vis. Comput. Graph. (2025)
24. Li, H., Wang, Y., Qu, H.: Where are we so far? understanding data storytelling tools from the perspective of human-AI collaboration. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (2024)
25. Lo Duca, A.: Become a Great Data Storyteller: Learn How You Can Drive Change with Data. John Wiley & Sons (2025)
26. Matuk, C., et al.: Data storytelling in the classroom. In: Proceedings of the 16th International ICLS 2022. International Society of the Learning Sciences (2022)
27. Matuk, C., Hurwich, T., Spiegel, A., Diamond, J.: How do teachers use comics to promote engagement, equity, and diversity in science classrooms? Res. Sci. Educ. **51**, 685–732 (2021)
28. Milesi, M.E., Martinez-Maldonado, R.: Data storytelling in learning analytics? a qualitative investigation into educators' perceptions of benefits and risks. In: Proceedings of the 14th LAK Conference (2024)
29. Mitchell, D., Sutherland, D.: What really works in special and inclusive education: using evidence-based teaching strategies. Routledge (2020)
30. Rose, D.H., Meyer, A.: Teaching every student in the digital age: Universal design for learning. ERIC (2002)
31. Roselli, D., Matthews, J., Talagala, N.: Managing bias in AI. In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 539–544. ACM (2019)
32. Russomanno, A., O'Modhrain, S., Gillespie, R.B., Rodger, M.W.: Refreshing refreshable braille displays. IEEE Trans. Haptics **8**(3), 287–297 (2015)
33. Saborío-Taylor, S., Rojas-Ramírez, F.: Universal design for learning and artificial intelligence in the digital era: fostering inclusion and autonomous learning. Int. J. Prof. Dev. Learners Learn. **6**(2) (2024)
34. Sarica, H.C., Yildirim, D.: Data storytelling in education: a systematic review. INTED2024 Proceedings, pp. 6990–6997 (2024)
35. Segel, E., Heer, J.: Narrative visualization: telling stories with data. IEEE Trans. Visual Comput. Graphics **16**(6), 1139–1148 (2010)
36. Shao, H., Martinez-Maldonado, R., Echeverria, V., Yan, L., Gasevic, D.: Data storytelling in data visualisation: does it enhance the efficiency and effectiveness of information retrieval and insights comprehension?. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–21 (2024)
37. Stahl, B.C., Eke, D.: The ethics of chatgpt-exploring the ethical issues of an emerging technology. Int. J. Inf. Manage. **74**, 102700 (2024)
38. Supplementary Material. https://github.com/alod83/AIED2025
39. Assistive technology. https://tinyurl.com/2jrvvw6f
40. The UDL Guidelines. https://udlguidelines.cast.org
41. UNESCO. https://tinyurl.com/3w5hydc6

42. Wei, Z., Qu, H., Xu, X.: Telling data stories with the hero's journey: Design guidance for creating data videos. IEEE Trans. Vis. Comput. Graph. (2024)
43. Yan, L., et al.: From data stories to dialogues: a randomised controlled trial of generative AI agents and data storytelling in enhancing data visualisation comprehension. arXiv preprint arXiv:2409.11645 (2024)
44. Yang, L., et al.: A design space for applying the freytag's pyramid structure to data stories. IEEE Trans. Visual Comput. Graphics **28**(1), 922–932 (2022)



Virtual Simulated Patients for Medical Interviews Using Large Language Models with a Self-refinement Mechanism to Suppress Excessive Responses

Naoki Shindo(✉) and Masaki Uto(✉)

The University of Electro-Communications, Chofu, Tokyo 182-8585, Japan
{shindo_naoki,uto}@ai.lab.uec.ac.jp

Abstract. In medical education, the medical interview is a fundamental skills that medical students need to learn. Practicing interviews with standardized patients (SPs), who simulate the behavior of actual patients, is known to be effective for developing medical interview skills. However, preparing qualified SPs who can act as patients in strict accordance with given clinical scenarios is both time- and cost-intensive. To address this issue, many studies have proposed virtual simulated patients (VSPs) utilizing large language models (LLMs). Conventional VSPs, however, often generate excessive responses, which are undesirable. Moreover, these methods typically rely on cloud-based LLMs, which raises significant concerns about data leakage and operational costs when used in certain educational and testing contexts. To overcome these challenges, this study proposes an approach to develop VSPs equipped with mechanisms to suppress excessive responses by utilizing an open-source LLM. Through experiments with actual data, we demonstrate that the proposed method significantly reduces excessive responses.

Keywords: Language models · medical education · simulated patient

1 Introduction

In medical education, the medical interview is a fundamental skill that medical students must acquire, which involves gathering information and building trust with patients [7]. Effective medical interviews require not only medical knowledge but also a multifaceted skill set, including empathy, active listening, and patient-centered communication [22]. Furthermore, the medical interview is included as one of the tasks in the objective structured clinical examination (OSCE) [24], a standardized test conducted across medical universities, which further emphasizes that medical students are expected to engage in learning and continuous skill development to achieve proficiency in medical interviews.

Simulated patients (SPs), who imitate the behavior of actual patients, play a crucial role in the learning and examination of medical interviews. For example,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 52–59, 2025.
https://doi.org/10.1007/978-3-031-99267-4_7

the medical interview examination in the OSCE assesses the ability of students by evaluating their conversations with SPs. Furthermore, repeatedly practicing interviews with SPs is effective for developing medical interview skills [11].

SPs are required to act as patients in strict accordance with clinical scenarios that specify details such as age, sex, family background, and specific symptoms[1]. However, preparing qualified SPs requires resource-intensive prior training [17], which makes it challenging to maintain a sufficient number of well-trained SPs. This limitation not only reduces opportunities for medical students to practice medical interviews with SPs but also increases the cost of managing the OSCE.

As an approach to addressing these challenges, numerous studies have proposed virtual simulated patients (VSPs) that replace human SPs using artificial intelligence (AI). Traditional methods have long relied on rule-based approaches, in which dialogue progresses by selecting responses from a manually predefined set of response rules [5,10,12,14–16,18,28]. However, such methods require the labor-intensive creation of response rule sets, incurring high costs and making it prohibitively expensive to develop VSPs capable of flexible interactions. To address these challenges, several studies have proposed VSPs based on large language models (LLMs) [1,4,6,19]. Typical LLM-based methods follow a straightforward approach, which prompts LLMs with detailed instructions to act as an SP within a specific scenario. Although these methods have succeeded in generating fluent and flexible responses, they often produce excessive responses, including scenario information unrelated to the given questions [21]. Such responses deviate from the purpose of a medical interview, where the focus is on eliciting required information from patients through effective questioning, and are thus unsuitable for VSPs.

A simple approach for suppressing excessive responses is to prompt LLM-based VSPs to respond as concisely as possible, focusing only on scenarios related to the given questions. However, our preliminary experiments found that this approach has a very limited effect. Another possible approach is to use the latest cloud-based LLMs with massive parameters because scaling the LLM generally enhances its ability to generate responses that better align with the given prompts [9]. However, relying on such cloud-based LLMs presents challenges from the following two perspectives when considering practical usage of VSPs.

1. **Data confidentiality**: In high-stakes examinations like the OSCE, examination content and interaction data must remain confidential. Cloud-based LLMs risk potential data breaches, making them less suitable.
2. **Operational cost**: Applications for examinations and daily learning require extensive and sustained usage by many students. Thus, the usage costs of cloud-based LLMs make large-scale and long-term deployment challenging.

For these reasons, VSPs utilizing local LLMs with limited computational requirements are desirable.

[1] An example of a clinical scenario provided to SPs can be found at https://github.com/AI-Behaviormetrics/LLM-VSPs/blob/main/Scenarios/scenario1.txt.

Table 1. Responses of proposed and baseline VSPs.

Question :	Have you noticed any changes in your weight recently?
Baseline :	Yes, I've gained about 3 kg *since April, when my back pain started. I've tried to maintain a balanced diet, but I've been eating more comfort food due to the pain and sedatives.*
Proposed :	Yes, I have gained about 3 kg.

Against this background, this study proposes a method for building VSPs equipped with a framework to suppress excessive responses, utilizing relatively small open-source LLMs that operate with limited computational resources. Specifically, we propose a novel approach that integrates two complementary components: *question category (QC)-aware answering* and the *self-refinement mechanism* to effectively suppress excessive responses without scaling LLMs.

This study experimentally demonstrates that the proposed method substantially reduces excessive responses using a relatively small open-source LLM as the base model. Table 1 shows examples of responses generated by the proposed and baseline methods for the same question, where italicized text highlights the excessiveness in baseline responses that is suppressed by the proposed method. Such improvements make the use of VSPs more practical.

2 Proposed Method

The proposed method functions as an LLM-based VSP through two steps. First, the LLM receives a question from a student in the doctor role and internally generates an initial response using the *QC-aware answering* framework. This response is then reviewed and revised through the *self-refinement mechanism* to produce the final output. The remaining part describes these two components.

Question Category-Aware Answering. The QC-aware answering mechanism helps the LLM recognize the categories of questions during response generation. According to [20], medical interview questions can be classified along three dimensions: *question type*, *context dependence*, and *symptom relevance*. Question type is classified as either open-ended or closed-ended, context dependence indicates whether the question is context-dependent or independent, and symptom relevance specifies whether the question is symptom-relevant or irrelevant. By combining these three dimensions, each question can be categorized into one of eight groups.

The granularity of responses required for each question category typically differs. For example, for the question "Did you measure your temperature?", which can be classified into the category {Closed Question, Context Independent,

Symptom Relevant}, the LLM is expected to generate a concise answer such as "yes" or "no", while avoiding other contextual information.

QC-aware answering is implemented using an implicit CoT approach [26, 27] combined with few-shot prompting [23,25]. The implicit CoT approach is a prompting technique that helps LLMs understand the reasoning process during response generation by explicitly including the detailed procedures required for reasoning in the prompt. Unlike in standard CoT prompting, in implicit CoT, the reasoning process itself is not explicitly output but is instead used internally to guide the response generation.

In our method, the prompt is designed to include explanations of the eight question categories, along with two example question-response pairs for each category and instructions on how to behave as a VSP[2].

Self-refinement Mechanism. The response generated through QC-aware answering is then reviewed and revised through the *self-refinement mechanism* to produce the final output. The self-refinement mechanism is an AI agent approach that has recently received significant attention [2]. In this study, it is implemented using a prompt that provides detailed instructions for evaluating the excessiveness of the given response[3].

Inspired by the method for LLM-based natural language evaluation [13], this prompt is designed to perform three sequential tasks: $< thinking >$, $< judgement >$, and $< feedback >$, with each step serving a specific purpose as follows:

1. $< thinking >$: The LLM gathers and organizes the information required to evaluate the response.
2. $< judgement >$: Based on the information from the thinking step, the LLM assesses the excessiveness of the response. At this stage, the LLM outputs either [*yes*] or [*no*] to indicate excessiveness.
3. $< feedback >$: If the judgement step results in [*no*], the LLM generates specific feedback to improve the response.

After receiving feedback, the LLM regenerates the response by incorporating the feedback. This process is repeated recursively until the LLM outputs [*yes*] in the $< judgement >$ step. Note that the same LLM is used for QC-aware answering, self-refinement, and response correction, meaning that our method requires only a single LLM. This study uses the Llama 3.1 8B model from Meta as the base open-source LLM.

3 Evaluation

We conducted experiments to evaluate the effectiveness of the proposed method.

[2] The prompt template is available at https://github.com/AI-Behaviormetrics/LLM-VSPs/blob/main/Prompts/Proposed/qc_aware_answering.txt.

[3] The prompt template is available at https://github.com/AI-Behaviormetrics/LLM-VSPs/blob/main/Prompts/Proposed/self_refine.txt.

For the experiments, three scenarios were created based on actual scenarios provided for SPs [3]. Each scenario includes detailed descriptions of symptoms related to the respective chief complaint, as well as additional contextual information such as age, sex, and family background. Additionally, a list of medical interview questions was prepared for each scenario. The chief complaints in the three scenarios are hernia, pyrexia, and hematemesis, with 79, 87, and 77 questions, respectively.

The responses generated by the proposed method for individual questions were evaluated using an automated technique known as "LLM-as-a-judge" [8], with OpenAI's GPT-4o serving as the judge LLM. The evaluation was conducted by prompting the judge LLM with a question–response pair, along with the entire scenario and detailed instructions for evaluation, following a dedicated rubric for each of the four criteria listed below[4].

- **Non-excessiveness**: Whether the response is concise and free of unnecessary information.
- **Relevance**: Whether the response is consistent with the provided scenario.
- **Instruction Robustness**: Whether the response faithfully adheres to the fundamental instruction of responding as a patient.
- **Naturalness**: Whether the response is natural in the context of a dialogue.

For each of these criteria, judgments were classified into two categories: [*yes*] and [*no*]. A [*yes*] judgment, indicating a positive evaluation, was encoded as 1, while a [*no*] judgment, indicating a negative evaluation, was encoded as 0. The average score across questions in each scenario was calculated for each evaluation criteria based on the encoded judgments.

For comparison, we ran the same experiments with the following methods:

- **Baseline method:** This is a conventional LLM-based approach [1,6] that incorporates a simple method to reduce excessiveness. It utilizes the Llama 3.1 8B model with a straightforward prompt consisting of a scenario and instructions directing the model to respond concisely in the role of the SPs.
- **Proposed method without QC-aware answering**: This method is implemented by removing the explanations of the question categories from the prompt. However, the two question–response examples for each category (16 examples in total) remain in the prompt.
- **Proposed method without self-refinement**: This method utilizes the response generated by the proposed method with QC-aware answering in the first trial, without executing the self-refinement process.

[4] The prompt templates for the LLM-as-a-judge are available at https://github.com/AI-Behaviormetrics/LLM-VSPs/tree/main/Prompts/LLM_Judge. Furthermore, we evaluated the agreement between GPT-4o and human judgments on 30 randomly selected questions from Scenario 1, yielding an overall average accuracy of 0.89 across four criteria, supporting the reliability of the LLM-as-a-judge approach.

Table 2 shows the experimental results, where the rows labeled "w/o QC", "w/o SR (self-refinement)" indicate the two variants of the proposed method explained above. Bold font indicates the best performance for each criterion and scenario.

Table 2. Experimental results.

		Non-excessiveness	Relevancy	Instruction Robustness	Naturalness
Scenario1	Baseline	0.101	0.886	0.987	**1.000**
	Proposed	**0.911**	0.873	**1.000**	**1.000**
	– w/o QC	0.759	0.797	**1.000**	**1.000**
	– w/o SR	0.228	**0.924**	**1.000**	**1.000**
Scenario2	Baseline	0.126	0.830	0.955	0.989
	Proposed	**0.885**	0.839	**1.000**	0.989
	– w/o QC	0.784	0.736	0.989	**1.000**
	– w/o SR	0.402	**0.874**	0.997	0.954
Scenario3	Baseline	0.128	**0.859**	0.949	0.987
	Proposed	**0.870**	0.805	**1.000**	**1.000**
	– w/o QC	0.792	0.857	0.987	0.974
	– w/o SR	0.506	0.857	**1.000**	0.974

First, we observe that the baseline method consistently yields very low non-excessiveness scores across all scenarios. This supports the claim that simply prompting an LLM to suppress excessiveness is ineffective, as described in Sect. 1. Next, comparing the baseline with the proposed method shows that the proposed method achieved substantially higher non-excessiveness scores across all scenarios, while the other evaluation criteria showed relatively high values for both methods. This suggests that the proposed method is highly effective in reducing excessiveness while maintaining high quality in other respects. Finally, comparing the proposed method with its variants reveals that omitting even one of the proposed components reduces the non-excessiveness score. This suggests that both proposed components contribute to reducing excessive responses.

To confirm whether there were significant differences in average performance among the methods for each evaluation criterion, we performed an analysis of variance. In all scenarios, the p-values for non-excessiveness were below the 5% significance level, while those for the other criteria exceeded 5%. A multiple comparison test further revealed significant differences in non-excessiveness scores across all method pairs. These results confirm that each proposed component improves non-excessiveness without harming performance on the other criteria.

4 Conclusion

In this study, we developed an LLM-based VSP for medical interviews using an open-source LLM that operates in a low-resource environment and incorporates a framework to reduce excessive responses. Through experiments with actual data, we demonstrated that the proposed method significantly reduces excessive responses. Future work includes more detailed experiments and methods to reduce response time caused by the recursive generation flow.

References

1. David, A.: Cook: creating virtual patients using large language models: scalable, global, and low cost. Med. Teach. **3**, 1–3 (2024)
2. Aman, M., et al.: Self-refine: iterative refinement with self-feedback. In: Advances in Neural Information Processing Systems. vol. 36, pp. 46534–4659 (2023)
3. Ban, N., Suzuki, T., Aomatsu, M., Saiki, T., Abe, K., Kuwabata, A.: An Easy-to-Understand Guide to Medical Interviews and Simulated Patients. The University of Nagoya Press [In Japanese] (2015)
4. Daniel, R., Kerstin, D.: Simulating diverse patient populations using patient vignettes and large language models. In: Proceedings of the First Workshop on Patient-Oriented Language Processing, pp. 20–25 (2024)
5. Dickerson, R., et al.: Evaluating a script-based approach for simulating patient-doctor interaction. In: International Conference of Human-Computer Interface Advances for Modeling and Simulation, pp. 79–84 (2005)
6. Eddi, G., Rashi, R., Ye Jean, P., Mehul, G.: OSCEai: personalized interactive learning for undergraduate medical education. Canadian Medical Education Journal (2024)
7. Harden, R.M.: What is an OSCE? Med. Teach. **10**(1), 19–22 (1988)
8. Jiawei, G., et al.: A survey on LLM-as-a-Judge (2024)
9. Kaplan, J., et al.: Scaling laws for neural language models. arXiv (2020)
10. Kleinheksel, A.: Transformative learning through virtual patient simulations: predicting critical student reflections. Clin. Simul. Nurs. **10**, 301–308 (2014)
11. Laschinger, S., et al.: Effectiveness of simulation on health profession students' knowledge, skills, confidence and satisfaction. Int. J. Evid. Based Healthc. **24**(278–302) (2008)
12. Liu, T., Luo, J., He, H., Zheng, J., Zhao, J., Li, K.: History-taking instruction for baccalaureate nursing students by virtual patient training: a retrospective study. Nurse Educ. Today **71**, 97–104 (2018)
13. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-Eval: NLG evaluation using GPT-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023)
14. Maicher, K.R., et al.: Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. Med. Teach. 279–285 (2022)
15. Parsons, T., Kenny, P., Rizzo, A.: Virtual human patients for training of clinical interview and communication skills. In: International Conference on Disability, Virtual Reality and Associated Technologies, p. 9 (2008)

16. Pereira, D., Falcão, F., Nunes, A., Santos, N., Costa, P., Pêgo, J.M.: Designing and building OSCEBot® for virtual OSCE-performance evaluation. Med. Educ. Online **28**(1), 16 (2023)
17. Rau, T., Fegert, J., Liebhardt, H.: How high are the personnel costs for OSCE? A financial report on management aspects. GMS J. Med. Educ. **28**(1), 16 (2011)
18. Robin, R., Benjamin, R., Maximilian, F.: Intent recognition in doctor-patient interviews. In: Twelfth Language Resources and Evaluation Conference, pp. 702–709. European Language Resources Association (2020)
19. Samuel, S., Rojin, Z., Carl, H., Eduardo, R., Jeffrey, J., Michael, M.: AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv (2024)
20. Shigeyuki, N.: Simulated patient (SP) and learning of medical communications. Jpn. Pharmacol. Ther. **38**(12), 1077–1088 (2010)
21. Shindo, N., Uto, M.: ChatGPT-based virtual standardized patient that amends overly detailed responses in objective structured clinical examinations. In: 25th International Conference on Artificial Intelligence in Education, pp. 263–269. Springer Nature Switzerland (2024)
22. Stewart, M.: Effective physician-patient communication and health outcomes: a review. CMAJ **10**, 1423–1433 (1995)
23. Tom, B., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901 (2020)
24. Uto, M., Tsuruta, J., Araki, K., Ueno, M.: Item response theory model highlighting rating scale of a rubric and raterârubric interaction in objective structured clinical examination. PLoS ONE **19**(9), 1–23 (2024)
25. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys (2020)
26. Yuntian, D., Kiran, P., Roland, F., Paul, S., Vishrav, C., Stuart, S.: Implicit chain of thought reasoning via knowledge distillation. arXiv (2023)
27. Yuntian, D., Yejin, C., Stuart, S.: From explicit CoT to implicit CoT: Learning to internalize CoT step by step. arXiv (2024)
28. Zini, J.E., Rizk, Y., Awad, M., Antoun, J.: Towards a deep learning question-answering specialized chatbot for objective structured clinical examinations. In: International Joint Conference on Neural Networks, pp. 1–9 (2019)



Playful Pathways to AI Literacy: Designing NLP Activities for Middle School AI Education

Alex Goslen[1], Claire Aguiar[2], Jessica Vandenberg[2], Veronica Cateté[2], Wookhee Min[2], and Bradford Mott[2(✉)]

[1] Elon University, Elon, NC 27244, USA
agoslen2@elon.edu
[2] North Carolina State University, Raleigh, NC 27695, USA
{cmaguiar,jvanden2,vmcatete,wmin,bwmott}@ncsu.edu

Abstract. As artificial intelligence (AI) becomes increasingly embedded in daily life, there is a growing need to develop engaging and effective educational materials to foster K-12 students' AI literacy. This paper presents the iterative design, development, and refinement of three playful learning activities focused on natural language processing (NLP), aimed at demystifying AI concepts for middle school students. The activities were implemented across two summer camps, with findings indicating increased engagement and improved understanding in the second camp. We also present design principles that emerged from this work to support the broader integration of AI literacy in educational settings.

Keywords: K-12 AI Education · Game-Based Learning · Middle School

1 Introduction

With rapid advancements in artificial intelligence (AI), its presence in teaching, learning, and everyday life has grown significantly. The ability to critically analyze and engage with AI is no longer just a technical skill—it is a fundamental literacy essential for all students [10,17]. However, despite the increasing need to promote AI literacy, many educators lack the resources and training needed to effectively incorporate AI concepts into their curricula, leaving students underprepared. Initiatives have started to address these challenges, such as AI4K12's Five Big Ideas in AI for K-12 education, which established foundational guidelines for teaching AI in both formal and informal settings [15].

Given the importance of fostering an AI-literate populace, it is essential to expand efforts that provide youth with early AI learning opportunities. Middle school is a critical period for shaping students' perceptions and dispositions toward STEM [4], making it an opportune time to introduce engaging AI learning experiences to them. A promising approach to enhancing engagement and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 60–67, 2025.
https://doi.org/10.1007/978-3-031-99267-4_8

comprehension is game-based learning [12,19]. Researchers have begun exploring the use of playful learning experiences to support AI education for K-12 students, particularly in areas such as search, planning, and machine learning [5,6,16].

Much of the previous work on playful AI literacy education has focused on the Big Ideas of *Learning* and *Representation & Reasoning* [5], while few activities have explored *Natural Interaction*. To address this gap, we have designed a set of playful learning activities centered on *Natural Interaction*. This Big Idea emphasizes that intelligent agents require a broad range of knowledge and capabilities to engage with humans naturally, often through the use of natural language. Natural Language Processing (NLP)—the subfield of artificial intelligence dedicated to enabling computers to generate and understand human language—plays a crucial role in enabling these types of rich human-machine interactions. In this paper, we present three NLP-focused learning activities designed specifically for middle school students. The first two activities are unplugged, or do not use a computer, and the third is a digital game-based learning activity.

2 Related Work

Game-based learning has been studied for many years, with computer games and simulations being utilized for educational purposes as early as the 1950s [18]. Contemporary research in this field has focused on refining core principles and advancing theoretical frameworks to deepen our understanding of the impact of game-based learning on student outcomes [1,8,13]. As research has evolved, studies have increasingly highlighted the potential of game-based approaches to enhance subject-specific learning, including computer science.

Researchers have started investigating how game-based learning can support AI literacy. Alam found that games offer a novel framework for classroom teaching of machine learning, artificial intelligence, and computational thinking [2]. Others recently introduced an unplugged approach to teaching ethical AI, where students role-play as an AI, a developer, or a tester, promoting discussions on fairness, transparency, and bias [7]. Druga et al. explored how conversational agents provide insight into how learners develop a conceptual understanding of human-AI interaction, demonstrating that playful engagement can deepen comprehension of how AI systems interpret and respond to human input [3]. Wang et al. explored teaching search algorithms through an engaging narrative-centered game [16], while Lim et al. introduced students to AI concepts related to facial recognition through an unplugged game-based activity [9]. These findings highlight the potential of game-based learning for AI literacy.

3 Natural Interaction Focused Learning Activities

To bring AI learning opportunities aligned with *Natural Interaction* to middle school students [15], we developed three playful activities that have been iteratively refined through multiple summer camp implementations.

3.1 SimilarityShuffle: A Card Embedding Game

SimilarityShuffle is an unplugged, game-based learning activity that introduces students to embeddings and similarity scores. The goal of the game is to help students understand distance-based similarity and provide practice calculating similarity scores within a vector space of embeddings. This activity was incorporated after the first camp to provide additional scaffolding, as students initially struggled with the concept of similarity scores. The activity utilizes a standard French-suited deck of cards, consisting of 13 ranks and 4 suits (diamonds, clubs, hearts, and spades). Students are presented with a *card embedding space*, where the deck of cards is organized into a structured 2D grid (Fig. 1). This allows students to visually assess how "close" (i.e., similar) cards are to one another and determine distances between them. For example, in this space, the Jack of Diamonds is three steps away from the Jack of Clubs using Manhattan distance.

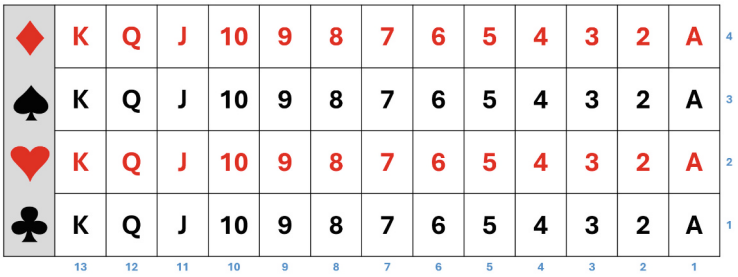


Fig. 1. Card embedding space for assessing distance in SimilarityShuffle.

To play the game, students are divided into groups of 4 to 8 players. One student acts as the dealer and distributes five cards to each player. The dealer then places a single card face-up in the center of the group. Students take turns playing the card most similar to the center card, aiming to win the round based on the *card embedding space*. The game continues until one player wins five rounds. The process of counting steps using the "card embedding" is designed to mimic how similarity scores are calculated in a word/sentence embedding space, reinforcing the concept of distance-based similarity in AI models.

3.2 SpongeWords: A Sentence Generation Challenge

SpongeWords is a playful, unplugged activity that introduces students to concepts related to language models. The activity simulates how language models generate sentences based on a given body of text and the probability of specific word sequences within that text. Students are provided with a paragraph of text summarizing the plot of an episode of the popular television series *SpongeBob SquarePants* [14]. They refer to this paragraph while completing the challenge.

Students are divided into groups of 4 to 8. They are then prompted to generate the next sentence of the paragraph by writing down the most probable word, with each student on the team taking turns selecting one word for the sentence. For example, if the body of text has a sentence like "SpongeBob and his best friend Patrick live in Bikini Bottom," and a student starts the next sentence with the word "SpongeBob" the next most probable word would be "and" because of the sentence above. This process continues until a full sentence is built.

In the first round of the activity, students are restricted to using words shown in the given paragraph, and in the second round, students are allowed to use any word in the English language. This structure is designed to teach students the idea of training language models on certain bodies of text. At the end of each round, teams read their sentences aloud, and the instructors select the winning sentence. This process teaches students about data annotation and how language models are often annotated by humans to foster natural interaction.

3.3 DialogueCrafter: A Chatbot Dialogue Design Challenge

DialogueCrafter is a digital game-based activity (Fig. 2), designed to help students explore NLP concepts by interacting with and creating dialogue for non-player characters (NPCs), while learning about AI in farming—a topic teachers identified as especially relevant given the central role of agriculture in their students' lives. After a brief overview, students interact with an NPC named Bonnie using natural language; however, she responds the same way to all questions. After several interactions with the NPC, students are asked to reflect on how her dialogue could be improved, fostering critical thinking about *Natural Interaction*.



Fig. 2. DialogueCrafter game-based learning environment.

The in-game activities then follow a Use-Modify-Create scaffolding progression [11]. In the *Use* phase, students view example question-answer (QA) pairs from a knowledge-base for an NPC and are prompted to ask additional questions. The NPC returns a similarity score and the answer to the closest matching question. In the *Modify* phase, students edit answers to questions and see their updates reflected through follow-up questions to the NPC. In the final *Create* phase, students generate new questions and answers for another NPC, observing how the NPC changes accordingly.

4 Implementations

The NLP activities were implemented across two summer camps involving 33 middle school students. Informed parental consent and student assent was obtained prior to participation, as approved by the university's ethics board.

Our first implementation occurred during summer 2023, as a week-long AI-focused camp with 10 middle school students from a rural community in Southeastern United States, all of whom had limited prior exposure to AI concepts. Participants included four girls and six boys, with 9 of 10 identifying as Black and the other identifying as Hispanic or Latino. One day of the camp was dedicated to *Natural Interaction*. The day started with an introduction to technologies that use NLP and challenges associated with understanding natural language. Afterwards, students engaged with the SpongeWords sentence generation activity and then played DialogueCrafter for roughly 30 min (Fig. 3). In this first camp, the NPCs in the game focused on different AI application topics instead of farming. At the end of the day, students completed an exit survey.



Fig. 3. Students playing DialogueCrafter during summer camp.

We observed that students experienced confusion with the similarity score presented in the game during interactions with NPCs, which prompted us to explore creating SimilarityShuffle as previously mentioned. In the exit survey, students demonstrated an understanding of voice assistants' ability to process natural language; however, students were unable to provide specific examples,

when asked about challenging scenarios for NLP technologies. This demonstrates that students struggled to understand the limitations of NLP. This feedback helped us to refine our introductory materials to include discussions about how language models can interpret the context of natural language and emphasize the limitations of technologies that use NLP. We also refined the DialogueCrafter activity to incorporate a new NPC, Bonnie, that has the same response for every inquiry with a reflection afterwards. This addition was designed to highlight limitations of NLP technologies when they do not work as expected.

Our second implementation occurred during summer 2024 with 23 middle school students participating. Of these students, 40% reported as female, 52% reported as male, and 8% reported other. Students also reported the following racial demographics: 26% Black, 26% White, 4% American Indian, 31% South Asian/Indian, 4% Middle Eastern, 8% Multiracial. In this implementation, we presented students with our refined introduction to NLP and all three game-based activities. Students played each activity for around 30 min. At the end of the day, students completed an exit survey.

When asked to select the correct definition of similarity score on the survey, 83% of students identified the correct definition. When asked to identify challenging scenarios for NLP, 79% were able to identify specific scenarios for the limitations of language models, which demonstrates a higher understanding than the previous camp. Additionally, several students identified SimilarityShuffle and SpongeWords as their favorite activities of the day.

5 Design Principles

Through refinements to our activities, we have identified key design principles for developing effective AI literacy materials. The first principle is to root activities in topics familiar to students. Due to the complex nature of AI, students with limited computer science experience may find it intimidating to learn. However, incorporating examples of familiar technologies, referencing popular television shows, and designing activities to align with students' everyday experiences appears to enhance engagement with the educational materials.

The second design principle is to integrate ethical discussions and real-world applications of AI. Beyond understanding AI mechanics, students should engage in conversations about the broader societal impacts of AI, such as bias, privacy, and fairness. Reflections can prompt students to consider how AI affects their daily lives and the ethical responsibilities of AI developers. By incorporating case studies or real-world examples, students can explore AI's benefits and limitations through meaningful, age-appropriate discussions.

The third key principle is to encourage iteration and experimentation within digital game-based activities. Allowing students to modify AI behaviors, test different inputs, and analyze outputs fosters a more hands-on, inquiry-driven approach. Iteration encourages problem solving and reinforces the idea that AI models are dynamic systems that can be refined over time. This principle mirrors computational thinking practices, where debugging and optimization are

essential parts of learning. Encouraging students to reflect on their process and refine their AI-based projects promotes a deeper understanding of AI systems.

6 Conclusion

This work introduced the design of three game-based activities designed to support middle school students in understanding *Natural Interaction*. Centered on natural language processing (NLP), the activities were implemented and improved across two summer camp implementations. Survey responses suggest increased student engagement and improved understanding of NLP concepts. These findings align with prior research on the benefits of game-based learning and point to the potential of such activities for formal classroom settings. Our work also surfaced a set of design principles to guide the development of effective, playful AI literacy experiences. Furthermore, we make the resources for these activities openly available to support educators.[1]

While promising, the findings reflect limitations related to small sample sizes and the informal nature of the camp settings. Broader implementation across diverse student populations and in-school contexts is needed to assess scalability and long-term impact. Additionally, tailoring activity content—such as offering a range of familiar media for the sentence generation task—may further boost relevance and engagement. While ethical considerations around AI were discussed during the implementations, they are not yet integrated directly into the gameplay itself; this represents an important area for future enhancement.

Looking ahead, we aim to explore the integration of large language models into student-facing tools, enabling richer conversations around AI's capabilities, limitations, and societal implications. Future iterations of these activities will incorporate ethical dimensions more directly and evaluate their classroom use.

Acknowledgments. This research was supported by the National Science Foundation under Grant DRL-2148680. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

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

References

1. Adipat, S., Laksana, K., Busayanon, K., Asawasowan, A., Adipat, B.: Engaging students in the learning process with game-based learning: the fundamental concepts. Int. J. Technol. Educ. **4**(3), 542–552 (2021)
2. Alam, A.: A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 69–74. IEEE (2022)

[1] http://go.ncsu.edu/aiplay.

3. Druga, S., Williams, R., Breazeal, C., Resnick, M.: "hey google is it ok if i eat you?" Initial explorations in child-agent interaction. In: Proceedings of the 2017 Conference on Interaction Design and Children, pp. 595–600 (2017)
4. Faber, M., Unfried, A., Wiebe, E.N., Corn, J., Townsend, L.W., Collins, T.L.: Student attitudes toward stem: the development of upper elementary school and middle/high school student surveys. In: 2013 ASEE Annual Conference & Exposition, pp. 23–1094 (2013)
5. Gennari, R., Melonio, A., Pellegrino, M.A., D'Angelo, M.: How to playfully teach AI to young learners: a systematic literature review. In: Proceedings of the 15th BI Annual Conference of the Italian SIGCHI Chapter, pp. 1–9 (2023)
6. Gupta, A., et al.: Supporting upper elementary students in learning AI concepts with story-driven game-based learning. In: AAAI Conference on Artificial Intelligence. vol. 38, pp. 23092–23100 (2024)
7. Henry, J., Hernalesteen, A., Collard, A.S.: Teaching artificial intelligence to k-12 through a role-playing game questioning the intelligence concept. KI-Künstliche Intelligenz **35**(2), 171–179 (2021)
8. Krath, J., Schürmann, L., Von Korflesch, H.F.: Revealing the theoretical basis of gamification: a systematic review and analysis of theory in research on gamification, serious games and game-based learning. Comput. Hum. Behav. **125**, 106963 (2021)
9. Lim, H., Min, W., Vandenberg, J., Cateté, V., Mott, B.: Unplugged k-12 AI learning: exploring representation and reasoning with a facial recognition game. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 23285–23293 (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 (2020)
11. Lytle, N., et al.: Use, modify, create: comparing computational thinking lesson progressions for stem classes. In: Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education, pp. 395–401 (2019)
12. McLaren, B.M., Nguyen, H.A.: Digital learning games in artificial intelligence in education (AIED): a review. Handbook of Artificial Intelligence in Education, pp. 440–484 (2023)
13. Plass, J.L., Homer, B.D., Kinzer, C.K.: Foundations of game-based learning. Educ. Psychol. **50**(4), 258–283 (2015)
14. Springer, A., Greenblatt, C., Williams, M.: SpongeBob SquarePants. Survival of the Idiots. Television series episode (2001). created by Stephen Hillenburg
15. Touretzky, D., Gardner-McCune, C., Martin, F., Seehorn, D.: Envisioning AI for k-12: What should every child know about AI? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9795–9799 (2019)
16. Wang, N., Greenwald, E., Montgomery, R., Leitner, M.: ARIN-561: an educational game for learning artificial intelligence for high-school students. In: International Conference on Artificial Intelligence in Education, pp. 528–531. Springer (2022)
17. Wang, N., Lester, J.: K-12 education in the age of AI: a call to action for k-12 AI literacy. Int. J. Artif. Intell. Educ. **33**(2), 228–232 (2023)
18. Wolfe, J., Crookall, D.: Developing a scientific knowledge of simulation/gaming. Simul. Gaming **29**(1), 7–19 (1998)
19. Wouters, P., Van Nimwegen, C., Van Oostendorp, H., Van Der Spek, E.D.: A meta-analysis of the cognitive and motivational effects of serious games. J. Educ. Psychol. **105**(2), 249 (2013)



Generating Diverse Difficulty Examples in Embedding Vector Spaces via Inverse Embedding

Yo Ehara(✉)

Tokyo Gakugei University, 4-1-1 Nukuikita-machi Koganei, Tokyo 1848501, Japan
ehara@u-gakugei.ac.jp

Abstract. The embedding of vector spaces, which numerically represent textual meaning, is a foundational technology in natural language processing (NLP). In the educational domain, difficulty scales are considered an essential aspect of meaning and are presumed to manifest as axes within these spaces. Although various difficulty scales, such as linguistic complexity and scientific content difficulty, exist, how these dimensions are embedded within vector spaces remains poorly understood. Recent advancements in inverse embedding techniques, originally developed for security applications, allow the reconstruction (generation) of textual examples from vector representations, even when the input vectors do not correspond to explicitly observed training data. In this study, we adapted the inverse embedding technology for educational applications and proposed a method for generating and evaluating textual examples corresponding to the candidate axes of cognitive difficulty in embedding spaces. By interpreting and assessing these difficulty-related axes using the generated examples, our approach provides insights into how different difficulty dimensions are encoded in vector spaces. This method has potential implications for personalized learning, automated readability assessments, and adaptive educational content generation.

Keywords: Inverse Embeddings · Generation · Difficulty Control

1 Introduction

The technique of "embedding," which derives vector representations that capture the meaning of texts, has become a foundational technology in modern natural language processing [16]. Conversely, the technique of "inverse embedding," which performs the exact opposite, reconstructing a text from its embedding vector, has recently gained attention [15]. Currently, inverse embedding is primarily studied for security purposes [15], for example, when investigating whether sensitive information such as personal names is leaked by reconstructing texts from embedding vectors exchanged in frameworks such as Retrieval-Augmented Generation (RAG) [17].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 68–76, 2025.
https://doi.org/10.1007/978-3-031-99267-4_9

In this study, we show that inverse embedding has potential applications beyond security, that is, controlling text difficulty based on intuitive operations in embedding spaces. We focus on *ordinal annotation*, where the order or ordered categories (e.g., difficulty [7,18], readability [4,12], and complexity [19]). The difficulty annotations are particularly significant in educational applications.

In educational applications, inverse embedding can be employed in the following ways. Figure 1 provides an overview of this study. *Extrapolation:* By generating examples corresponding to extended embedding vectors of very difficult texts, we can generate texts that are more difficult than those of the given two levels. In Fig. 1 (a), extrapolation is used to identify the coordinates of more difficult examples. Inverse embedding is then applied to these coordinates to generate text examples. *Interpolation:* For example, when both a simple text and a (very) difficult text are available, linear interpolation can be used to identify intermediate points. This enables the generation of intermediate text examples that illustrate the progression from simple to difficult texts, which helps determine a suitable learning sequence. In Fig. 1 (b), interpolation is used to derive intermediate examples, and inverse embedding is again applied to the coordinates to obtain text examples.

In machine learning, the use of the difference vector between the average (mean) vectors of two levels is theoretically justified. Calculating this vector corresponds to obtaining the solution of the relaxation of an optimization problem called pairwise linear discriminant analysis [11], which seeks the projection direction that best preserves the order and magnitude relations of points from both levels when projected onto a vector. The contributions of this paper are as follows: 1) We show that inverse embeddings enable us to generate texts by controlling their difficulty; 2) unlike previous HCI applications regarding text difficulty [9,13], our method is capable of handling "diverse" difficulties and enables users to choose which difficulty to control; and 3) by using carefully annotated text difficulty datasets, we evaluate our method.

2 Text Difficulty Dataset

In this study, we used *CEFR-SP*, a dataset annotated with Common European Framework of Reference for Languages (CEFR) levels for English sentences aimed at second language learners [1]. CEFR-SP consists of two sub-datasets: Wiki-Auto and SCORE. For this research, we used the SCORE subset.

In the CEFR framework, the levels range from simpler to more difficult as follows: A1, A2, B1, and B2. Levels with higher letters (B vs. A) and numbers (2 vs. 1) indicate greater difficulty. In CEFR-SP, each sentence is annotated by two annotators. To simplify the analysis, we used only sentences in which both annotators agreed at that level.

Reconstruction Experiments with Inverse Embedding: First, we examined how well inverse embedding could reconstruct original sentences from their embedding vectors. Throughout this study, we used the code provided by [15] (https://github.com/jxmorris12/vec2text). Inverse embedding typically requires training to determine appropriate text outputs for the given embedding vectors.

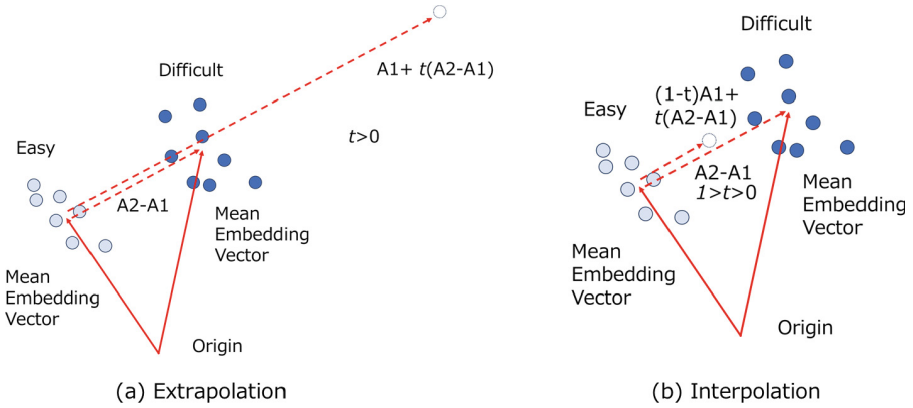


Fig. 1. Overview.

Table 1. Examples of how accurately inverse embedding can reconstruct vectors.

Sent. No.	Original	Recovered by Inverse Embedding
1	She had a beautiful necklace around her neck.	She had a beautiful necklace around her neck.
2	Larry has a shark 's tooth.	Larry's tooth is a shark tooth
3	Harrison got very excited when his parents gave him a car.	Harrison was very excited when his parents gave him a car he got a very excited about.
4	It 's hard to tell if these are my socks or his.	It's hard to tell if my socks are my socks or my socks.

Table 2. Average vectors for each level were calculated, and inverse embedding was applied (without normalization).

A1	No one was wearing it.
A2	I wish my sister was in the closet.
B1	She felt that the girls were very insecure when she went to the closet.
B2	She sucks her nieces urine from the nannies.

The code used in this study leverages a pretrained inverse embedding model based on OpenAI's *text-embedding-ada-002* sentence embedding vectors. *text-embedding-ada-002* generates 1,500-dimensional vectors that are normalized to a unit norm. Table 1 shows examples of sentence reconstruction. Sentence numbers are assigned within the table for ease of reference.

In Example 1, the reconstructed sentence matches the original almost perfectly, except for minor differences in the placement of the period. This high accuracy was likely obtained because the embedded text was relatively simple. Semantic considerations in inverse embedding are evident in cases such as Sentence number 2 in Table 1. In this example, the original sentence uses "has" and the reconstructed sentence uses "is," but it maintains almost the same meaning. This demonstrates that the inverse embedding considers the semantic content of the embedding vector. However, when dealing with longer or more complex sentences, the original text may not always be accurately reconstructed. For

Table 3. Cosine similarity between direction vectors representing difficulty. For example, A2-A1 represents the vector obtained by subtracting the average embedding vector of A1-level sentences from that of A2-level sentences in CEFR.

–	A2-A1	B1-A1	B1-A2	B2-A1	B2-A2	B2-B1
A2-A1	1.000	0.920	0.198	0.648	0.046	−0.031
B1-A1	0.920	1.000	0.566	0.685	0.157	−0.061
B1-A2	0.198	0.566	1.000	0.349	0.298	−0.088
B2-A1	0.648	0.685	0.349	1.000	0.790	0.685
B2-A2	0.046	0.157	0.298	0.790	1.000	0.925
B2-B1	−0.031	−0.061	−0.088	0.685	0.925	1.000

Table 4. Quantitative Evaluation of Extrapolated Vectors by using 10-nearest neighbors.

v+0.5(B2-A1)	1.9
v+0.9(B2-A1)	1.9
v+1.0(B2-A1)	2.3
v+1.1(B2-A1)	2.5
v+1.3(B2-A1)	2.8
v+1.5(B2-A1)	3.0
v+1.7(B2-A1)	3.1
v+1.9(B2-A1)	3.2
v+2.8(B2-A1)	3.4

instance, in example 3, although the reconstructed sentence is semantically similar to the original, it includes redundant phrases like "he got a very excited about." Additionally, the inclusion of "a" results in a grammatically incorrect sentence. This indicates that inverse embedding does not always yield a precise reconstruction of the original text. In some cases, the meaning of the reconstructed text diverges from that of the original. In example 4, the original sentence includes the word "his," but this word is missing in the reconstructed version, resulting in a failure to fully recover the original meaning.

Inverse Embedding of Average Vectors for Each Level: Next, we calculated the sentence embedding vectors for all the examples at each difficulty level, computed their simple average vectors, and applied inverse embedding to determine the sentences represented by the average vectors. The results are listed in Table 2. From Table 2, it is evident that as the level becomes more difficult, more complex words are used and the sentences become longer. This indicates that even when inverse embedding is applied, the average embedding vectors for each level reliably represent sentences at that level.

3 Experiments

Direction Vectors Representing Various Levels of Difficulty: As confirmed in Sect. 2, the "average sentence examples" calculated using the inverse embedding from the average embedding vectors for each level appear to be qualitatively appropriate. We consider the vector obtained by subtracting the average embedding vector for a simpler level from the average embedding vector for a more difficult level: This vector can be interpreted as the "direction" in which a sentence embedding vector should move, relative to the simpler sentence embedding vector, to become more difficult. Considering the difficulty of being a multi-scale attribute, there are multiple pairs of levels with different difficulties. For instance, in the CEFR-SP, there are four levels: A1, A2, B1, and B2. The number of methods for selecting two levels, one more difficult and one simpler, is $C(4,2) = 4 \times 3/2 = 6$, where C is the binomial coefficient. This implies that in

Table 5. Results of adding the vector difference between the average vector of B2-level sentences and the average vector of A1-level sentences to a simple sentence, and then applying inverse embedding. Sentences gradually become more complex, but after 2.0, the word "necklace" disappears, and the original structure is nearly lost.

v	She had a beautiful necklace around her neck.
v+0.4(B2-A1)	She had a beautiful necklace around her neck around her neck.
v+1.0(B2-A1)	She had a beautiful necklace around her neck adorned with a beautiful pearl necklace.
v+1.1(B2-A1)	Her necklace was very beautiful around her neck adorned with a small pearl necklace.
v+1.3(B2-A1)	Her necklace was very beautiful around her neck adorned with a small necropolis necklace.
v+1.7(B2-A1)	Her necklace was admired greatly by a young female nurse from a very fine necropolis etch.
v+1.8(B2-A1)	Her necklace was admired greatly by a young female narwhal from the famous jewelry store.
v+2.0(B2-A1)	Her narwhal necklaces were greatly admired by the female researcher from a small history center.
v+3.0(B2-A1)	The history of narwhal bones from a female sampler is greatly admired by female samplers.
v+5.0(B2-A1)	The history of forensics graduates from acorns can be greatly delved into by a modern forensic scientist.
v+6.0(B2-A1)	The history of forensics graduates from acorn schools is greatly delved into by forensic scientist.
v+9.0(B2-A1)	History of forensics graduates from avowed universities is compiled from the work of avowed forensic scientist.

the sentence embedding vector space, there are six distinct "directions for making sentences more difficult." We calculated the cosine similarities between these "directions for making sentences more difficult," which are shown in Table 3.

In Table 3, if the "direction for making sentences more difficult" is uniquely defined in the embedding vector space, all the differences between the average vectors for different levels would point in nearly the same direction, resulting in cosine similarities close to 1.0. However, this was not the case in the present study. For example, the "direction for making sentences more difficult" obtained by subtracting the average vector for A1 from that for B2 has a cosine similarity of 0.349 with the "direction for making sentences more difficult" derived from the difference between B1 and A1, indicating low correlation. This suggests that the "directions for making sentences more difficult" are diverse.

Qualitative Experiments: The results of a qualitative experiment in which difference vectors for each level were added to a simple sentence and inverse embedding was applied to make the sentences more difficult are presented in Table 5. In Table 5, v denotes the embedding vector of the original sentence. For simplic-

ity, we consider the average embedding vectors for each level and the examples at each level to be equivalent. Thus, $v + 0.4(B2 - A1)$ represents the result of adding 0.4 times the difference vector (average vector of B2-level examples − average vector of A1-level examples) to the embedding vector of the original sentence, followed by inverse embedding. As mentioned earlier, the difference vectors between the average embedding vectors for each level can be interpreted as "directions for making sentences more difficult." By increasing the coefficient (e.g., 0.4) or adding more difference vectors, the sentence becomes progressively more difficult. Note that the embedding space is normalized such that all sentence vectors have a norm of 1; therefore, the resulting vector is normalized to have a norm of 1. The Euclidean norm was used for normalization.

The results are listed in Table 5. Qualitatively, as the coefficient of the difference vector increases, the sentence becomes more difficult. Interestingly, up to a coefficient of approximately 2.0, the word "necklace" from the original sentence remains, but as the coefficient increases further, the generated sentences use difficult words unrelated to the original text. For example, at a coefficient of 1.8 or higher, the word "narwhal" appears, which refers to a "unicorn whale." According to Weblio's dictionary of learning levels, "Narwhal" is a word expected for TOEIC scores of 950 or higher (https://ejje.weblio.jp/content/narwhal), and in this test the theoretically highest score is 990.

Quantitative Experiment of Difficulty: We evaluated the difficulty of the extrapolated vectors using the mean values of the 10-nearest neighbors (NNs) in the embedding space, where A1, A2, B1, and B2 are 1.0, 2.0, 3.0, and 4.0, respectively, in Table 4. This indicates that larger values are more difficult. It can be observed that the extrapolated vectors are more difficult.

Quantitative Experiments using MMLU: We conducted similar experiments in Science, Technology, Engineering, and Mathematics (STEM) using the MMLU [8] dataset. We annotated 150 high school physics questions at the Japanese high school grade level. We divided the data into 100 training and 50 test data and used **multilingual-e5-small** embedding in HuggingFace. The results were v+0.5d: 1.4, v+0.6d: 1.6, v+0.7d: 1.7, and v+0.8d: 1.9, where v is the mean vector of the Grade 1 questions, d is the difference vector between Grades 1 and 2, and the values are the mean grades of the 10-NN test vectors. We can see that the extrapolated vectors are more difficult to use in physics.

4 Related Work and Discussion

We believe that the theoretical basis for successfully adjusting the difficulty via extrapolation of the average embedding vectors lies in the fact that the difference between the average embeddings corresponds to the solution of pairwise LDA [11], which is one of the simplest models of ordinal regression, such as Ranking SVM [2]. In the field of AI in education, some studies, such as [10,20], have addressed the problem of difficulty control in question generation. However, these studies involved inverse embeddings, which provide a cognitive understanding of the difficulties in embedding spaces. Without understanding

the use of embedding spaces, it is difficult to find the diversity of being difficult. This aspect is important in education because this field inevitably deals with diverse difficulties when involving diverse students, as it aims to be inclusive. Moreover, these studies focused on text generation in domain-specific problems, whereas we addressed the difficulty in controlling general text-generation settings using embedding spaces and inverse embeddings. In natural language processing, some studies, such as [3,14] address controlling the difficulty of generated texts. However, because these methods do not use textual embeddings, they are not intuitive, and do not address the diversity of difficulties, which is important for inclusive education. The author previously addressed the generation of educational content other than questions [5,6], whereas these studies do not involve embedding techniques or inverse embedding.

4.1 Limitations and Ethical Considerations

Notably, our method is specifically designed to control the difficulty of the generated texts; however, it does not guarantee the production of meaning-preserving paraphrases. This behavior is based on design; only nearby vectors encode semantically similar examples. Hence, it is reasonable to assume that an easy sentence and its distantly extrapolated difficult sentences are semantically dissimilar. Hence, our method is not designed for text-simplification tasks that require the retention of meaning. Our method has ethical risks if we use sentences generated in education as is. Although our method regulates textual difficulty, infrequent tokens, such as profanities or obscene expressions, can inadvertently emerge as advanced vocabulary due to their low corpus frequency. We conducted experiments on the ESL and MMLU benchmarks, and further evaluation is needed to demonstrate that our method is generally applicable to most types of educational content.

5 Conclusion

We demonstrated that sentence difficulty can be controlled using inverse embedding techniques and actual datasets of sentence difficulty via qualitative and quantitative experiments. We also showed that the "directions for making sentences more difficult" are diverse. Future studies will involve a more detailed analysis of text generation in the STEM field.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number 22K12287 and by JST, PRESTO Grant Number JPMJPR2363.

References

1. Arase, Y., Uchida, S., Kajiwara, T.: CEFR-based sentence difficulty annotation and assessment. In: Proceedings of EMNLP, pp. 6206–6219 (2022)
2. Cao, Y., Xu, J., Liu, T.Y., Li, H., Huang, Y., Hon, H.W.: Adapting ranking SVM to document retrieval. In: Proceedings of SIGIR, pp. 186–193 (2006)
3. Cheng, Y., et al.: Guiding the growth: difficulty-controllable question generation through step-by-step rewriting. In: Proceedings of ACL-IJCNLP, pp. 5968–5978 (2021)
4. Crossley, S., Heintz, A., Choi, J.S., Batchelor, J., Karimi, M., Malatinszky, A.: A large-scaled corpus for assessing text readability. Behav. Res. Methods 1–17 (2022). https://doi.org/10.3758/s13428-022-01802-x
5. Ehara, Y.: Innovative software to efficiently learn English through extensive reading and personalized vocabulary acquisition. In: Proceedings of AIED, pp. 187–192. Cham (2023)
6. Ehara, Y.: Measuring similarity between manual course concepts and ChatGPT-generated course concepts. In: Proceedings of EDM, pp. 474–476 (2023). https://doi.org/10.5281/zenodo.8115758
7. Hashimoto, B., Egbert, J.: More than frequency? Exploring predictors of word difficulty for second language learners. Language Learning (2019). https://api.semanticscholar.org/CorpusID:181868242
8. Hendrycks, D., et al.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
9. Hösel, C., Labude, S., Christoph, E., Roschke, C., Baumgart, M., Ritter, M.: Participation in the intelligibility testing of easy-to-read texts: development of an interface for people with learning difficulties. In: HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23, pp. 528–534. Springer (2021)
10. Jiao, Y., Shridhar, K., Cui, P., Zhou, W., Sachan, M.: Automatic educational question generation with difficulty level controls. In: International Conference on Artificial Intelligence in Education, pp. 476–488. Springer (2023)
11. Kamiya, H., Takemura, A.: On rankings generated by pairwise linear discriminant analysis of$m$Populations. J. Multivar. Anal. **61**(1), 1–28 (1997). https://doi.org/10.1006/jmva.1997.1662
12. Kincaid, P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)
13. Kojic, T., Vergari, M., Möller, S., Voigt-Antons, J.N.: Assessing user experience of text readability with eye tracking in virtual reality. In: International Conference on Human-Computer Interaction, pp. 199–211. Springer (2022)
14. Li, K., Zhang, Y.: Planning first, question second: an LLM-guided method for controllable question generation. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 4715–4729 (2024)
15. Morris, J.X., Zhao, W., Chiu, J.T., Shmatikov, V., Rush, A.M.: Language model inversion (2023)
16. Reimers, N.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019)
17. Salemi, A., Zamani, H.: Evaluating retrieval quality in retrieval-augmented generation. In: Proceedings of SIGIR, pp. 2395–2400 (2024)

18. Settles, B., LaFlair, G.T., Hagiwara, M.: Machine learning–driven language assessment. Trans. Assoc. Comput. Linguist. **8**, 247–263 (2020). https://doi.org/10.1162/tacl_a_00310
19. Smădu, R.A., Ion, D.G., Cercel, D.C., Pop, F., Cercel, M.C.: Investigating large language models for complex word identification in multilingual and multidomain setups. In: Proceedings of EMNLP, pp. 16764–16800 (2024)
20. Tomikawa, Y., Uto, M.: Difficulty-controllable multiple-choice question generation for reading comprehension using item response theory. In: International Conference on Artificial Intelligence in Education, pp. 312–320. Springer (2024)

Knowledge Tracing with A Temporal Hypergraph Memory Network

Mehrnoush Mohammadi[1(✉)], Kamal Berahmand[2], Shazia Sadiq[1], and Hassan Khosravi[1]

[1] The University of Queensland, Brisbane, Australia
{m.mohammadibolbanabad,h.khosravi}@uq.edu.au, shazia@eecs.uq.edu.au
[2] Queensland University of Technology, Brisbane, Australia
kamal.berahmand@hdr.qut.edu.au

Abstract. Knowledge tracing models students' evolving knowledge states to predict their future performance based on the question-answer interactions. Memory-based models have advanced KT by extracting latent knowledge concepts of questions and modeling students' knowledge using a dual memory mechanism, but fail to capture how concepts behave differently when appearing in various combinations. We propose the Temporal Hypergraph Memory Network (THMN), which addresses this limitation through three key innovations: (1) modeling concept interactions via temporal hypergraph structures with bidirectional message-passing between concepts and questions, (2) integrating these rich concept representations with knowledge states through attention-based fusion, and (3) employing an adaptive scaling mechanism that recognizes concepts are better learned when practiced in diverse contexts. Experiments on four benchmark datasets show THMN outperforms seven state-of-the-art methods, highlighting the importance of modeling concept interactions.

Keywords: Knowledge Tracing · Memory Networks · Concept Interactions · Temporal Hypergraphs

1 Introduction

Knowledge Tracing (KT) is a foundational task in educational systems, enabling dynamic modeling and prediction of students' knowledge states across various concepts over time [15]. By providing real-time insights into learning progress, KT empowers educational stakeholders, from educators and administrators to intelligent tutoring systems and learning platforms, to make informed data-driven decisions, optimize learning paths, and address individual needs effectively [2].

The field has evolved from early probabilistic approaches like Bayesian Knowledge Tracing [10,21] and Item Response Theory [3] to deep learning models [16] broadly categorized as sequential [12,14], memory-augmented

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 77–85, 2025.
https://doi.org/10.1007/978-3-031-99267-4_10

[1], graph-based [13,20], and attention-based [6,9] models. Among them, memory-augmented models, particularly Dynamic Key-Value Memory Networks (DKVMNs) [22] have demonstrated superior performance by addressing the fundamental limitations of earlier models. DKVMNs employ a dual architecture with a static key matrix for concept representation and a dynamic value matrix for tracking student knowledge, offering enhanced interpretability, better handling long-term dependencies, and more precise concept-level knowledge modeling [1].

Despite these advances, memory-based models consider concepts as independent entities, failing to capture how concepts behave differently when appearing in various combinations. In educational contexts, concepts rarely exist in isolation, they interact in complex ways, with some combinations being more challenging for students. To address this limitation, we propose the Temporal Hypergraph Memory Network (THMN), a novel framework that integrates temporal hypergraph structures with memory-based knowledge tracing. THMN captures the dynamic nature of concept interactions through three key innovations: (1) a temporal hypergraph structure that models higher-order relationships between concepts in questions, (2) a bidirectional message-passing mechanism between concepts and questions that learns how concepts behave in different combinations, and (3) an adaptive scaling mechanism that adjusts knowledge updates based on concept practice diversity. Our key contributions are as follows:

- We introduce a temporal hypergraph structure with bidirectional message-passing to model dynamic concept interactions and learn rich representations that reflect how concepts behave differently in various combinations.
- We integrate this representation with a memory-based model through an attention-based fusion mechanism that dynamically weighs and combines these complementary information sources.
- We propose an adaptive scaling mechanism based on practice diversity, yielding more stable knowledge states for concepts practiced in varied contexts.
- Our model reveals the dynamic interplay between concepts and questions, enhancing the interpretability and practical value of knowledge tracing.
- Our experiments on four benchmark datasets demonstrate the superiority of THMN in both prediction accuracy and knowledge evolution modeling.

2 Preliminaries and Problem Formulation

KT aims to predict student's performance by modeling their knowledge state evolution through question-answer interactions. Given the concept set $C = \{c_1, ..., c_N\}$, question set $Q = \{q_1, ..., q_{|Q|}\}$, where each question may involve multiple concepts (high-order relationships), and the student's interaction sequence up to $t-1$, $X = \{x^1, ..., x^{t-1}\}$, where $x^t = (q^t, y^t)$ pairs the question tag $q^t \in Q$ and the response correctness $y^t \in \{0, 1\}$, KT aims to predict $p(y^t = 1|q^t, X)$. The following definitions support our model.

Definition 1 ***Hypergraph.*** *A hypergraph is defined as* $G = (V, E, W)$*, where* $V = \{v_1, \ldots, v_n\}$ *is the set of nodes,* $E = \{e_1, \ldots, e_m\}$ *is the set of hyperedges, and* $W \in \mathbb{R}^{m \times m}$ *is a diagonal matrix of the weights of the hyperedges [4]. In* G*, each* $e_i \subseteq V$ *can connect multiple nodes.* G *can be represented by an incidence matrix* $H \in \mathbb{R}^{n \times m}$*, where* $h(v, e) = 1$ *if* $v \in e$*, and* 0 *otherwise.*

Definition 2 ***Temporal Hypergraph.*** *A temporal hypergraph* $\mathcal{G}$ *[5] is a sequence* $\{G^1, ..., G^t\}$ *where* $G^t = (V, E^t, W^t)$*. The node set* V *remains fixed while* $E^t = E^{t-1} \cup e^t$ *evolves by adding new hyperedges.*

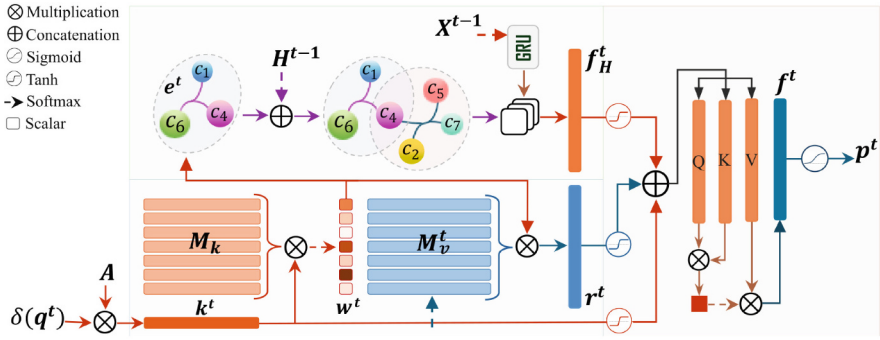


Fig. 1. THMN architecture for performance prediction at time step t, integrating student knowledge states and concept interaction patterns for accurate response modeling.

3 Proposed Method

In this section, we introduce THMN, a novel temporal hypergraph memory-based model for KT.

THMN involves three key structures: (1) a key-value memory pair $\langle M_k, M_v^t \rangle$, where static key matrix $M_k \in \mathbb{R}^{N \times d_k}$ stores latent concept representations and dynamic value matrix $M_v^t \in \mathbb{R}^{N \times d_v}$ tracks evolving student knowledge states on concepts through question-answering interactions [22], (2) a temporal hypergraph structure $H^t \in \mathbb{R}^{N \times |E^t|}$ that captures how concepts combine in questions over time, and (3) a dynamic concept interaction matrix $X_V^t \in \mathbb{R}^{N \times d_k}$ models concept behaviors in different combinations. THMN processes information through four main components: (1) Correlation Weight Computation, (2) Temporal Information Extraction, (3) Information Fusion and Answer prediction, and (4) Temporal Pattern Evolution. Figure 1 illustrated the THMN process.

3.1 Correlation Weight Computation

This section aims to extract the weights of concepts related to each question. Given the question tag $q^t \in Q$, its unique continuous representation $k^t \in \mathbb{R}^{d_k}$

is obtained by transforming its corresponding one-hot vector, $\delta(q^t)$ with lengh $|Q|$, through an embedding matrix $A \in \mathbb{R}^{|Q| \times d_k}$ as: $k^t = \delta(q^t)A$. Then, the correlation weight $w^t \in \mathbb{R}^N$ is calculated as: $w^t = \text{softmax}(M_k k^t)$, where w_i^t is the correlation weight between q^t and c_i.

3.2 Temporal Information Extraction

In this section, we extract two types of temporal information: knowledge states and concept interactions, providing complementary views of learning progress.

Knowledge State Retrieval. Given the correlation weights w^t, the read content $r^t \in \mathbb{R}^{d_v}$ is retrieved from M^v by $r^t = (w^t)^T M_v^t$, where r^t represents the student's knowledge state on q^t across its related concepts' knowledge states [22].

Concepts Interactions Processing. As students interact with questions, the higher-order relationships between concepts continuously evolve. We use temporal hypergraph structures to model concepts behaviors in different questions and extract their collective interaction patterns.

Hypergraph Construction. For new question q^t, we create a hyperedge $e^t \in \mathbb{R}^N$ by applying a relative thresholding mechanism to w^t as: $e_i^t = w_i^t$ if $w_i^t \geq \tau \cdot w_{\max}^t$, otherwise $e_i^t = 0$. The temporal hypergraph, $H^t \in \mathbb{R}^{N,*}$, is then updated as: $H^t = [H^{t-1}; e^t]$ if q^t is new. This approach preserves only significant concept relationships while adapting to different weight distributions.

Learning Concept Behaviors. To model concept behaviors in question-based combinations, H^t, we adopt the Hypergraph Networks with Hyperedge Neurons (HNHN) framework [8]. Suppose $X_E^t \in \mathbb{R}^{|E^t| \times d_k}$ is the representation matrix of hyperedges. Our model learns through a bidirectional message passing: (1) Concepts to questions: $X_E^t = \text{ReLU}(D_E^{-1}(H^t)^T D_V X_V^t W_E + b_E)$, and (2) Questions to concepts: $X_V^t = \text{Sigmoid}(D_V^{-1} H^t D_E X_E^t W_V + b_V)$, D_V and D_E are diagonal degree matrices. Finally, we compute the hypergraph summary f_H^t as:

$$f_H^t = (\tilde{e}^t)^T X_V^t \tag{1}$$

$\tilde{e}^t$ is the binarized vector of e^t. f_H^t captures behaviors of concepts when integrated into q^t, reflecting their learned interaction patterns from previous questions.

3.3 Information Fusion and Answer Prediction

To predict student performance, we integrate three information sources using scaled dot-product attention: k^t (question characteristics), r^t (student mastery), and f_H^t (Concepts' behavior in the interaction within q^t). After transforming

these vectors through fully connected layers with tanh activations and concatenating them as $C^t = [\tilde{k}^t; \tilde{r}^t; \tilde{f}_H^t]$, we compute: $f^t = \text{Softmax}\left(QK^T/\sqrt{d}\right)V$, where Q, K, and V vectors are derived from C^t, and d is the feature space dimension. Finally, the probability of correctly answering q^t is predicted as:

$$p^t = \text{Sigmoid}(W_p f^t + b_p) \tag{2}$$

This attention mechanism dynamically focuses on the most relevant aspects of each information source for accurate prediction.

3.4 Temporal Pattern Evolution

Knowledge State Evolution. This process sequentially updates the student's knowledge state on concepts M_v^t based on their response to q^t. We improve this process with an adaptive scaling mechanism based on the practice diversity.

For each interaction $x^t = (q^t, y^t)$, we compute erase and add vectors: $\varepsilon^t = Sigmoid(W_e^T \cdot v^t + b_e)$ and $a^t = Tanh(W_a^T.v^t + b_a)$, where v^t is the vector of knowledge growth, derived from x^t through an embedding matrix $B \in \mathbb{R}^{2|Q| \times d_v}$. Our key innovation is a diversity-based scaling factor $\beta_i^t = Sigma(1 - d_i^t/|E^t|)$ for each concept c_i, where d_i^t is the concept's degree in H^t and $|E^t|$ is the total number of hyperedges. The knowledge state is updated as:

$$M_v^{t+1} = M_v^t[1 - (w^t \odot \beta^t)(\varepsilon^t)^T]^T, \text{ then, } M_v^{t+1} = M_v^{t+1} + (w^t \odot \beta^t)(a^t)^T \tag{3}$$

This scaling mechanism reflects how concepts become better learned and memorized through diverse practice, leading to more stable knowledge states.

Concepts Interaction Evolution. We use a Gated Recurrent Unit (GRU) to capture how concepts behave in various combinations over time within the sequences as: $X_V^{t+1} = GRU(X_V^t)$. The GRU's gate mechanisms maintain important long-term dependencies while filtering irrelevant information, ensuring concept representations evolve meaningfully through the learning sequence.

3.5 Model Training and Optimization

We train THMN by minimizing the cross-entropy loss between predicted probabilities p^t and true responses y^t: $\mathcal{L} = -\sum_{t=1}^{T}(y^t \log(p^t) + (1 - y^t)\log(1 - p^t))$. The model is optimized using SGD with momentum 0.6 and gradient clipping at 50.0, jointly learning embedding and main matrices, and all other parameters.

4 Experiments

In this section, we empirically evaluate the effectiveness of THMN by addressing two research questions: **RQ1.** How does THMN perform compared to state-of-the-art KT models in prediction accuracy across benchmark datasets? and **RQ2.** How effectively does THMN model the concept of knowledge evolution?

4.1 Experimental Setting

Dataset and Baselines. We evaluate THMN on four benchmark datasets, Statics2011, Kddcup2010 [1], Synthetic-5 [14], and ASSISTments2009[1], by comparing its effectiveness with two groups of deep learning-based models. Three foundational KT models: DKT [14], DKVMN, and GKT [13], and four recent state-of-the-art methods: SKT [17], DGEKT [7], HawkesKT [18], and KTMFF [19].

Implementation Details. We randomly split each dataset into training (70%), validation (20% of training), and test sets (30%) by student ID. THMN was implemented by MXNet framework with Xavier initialization for layers. We set τ to 0.6, memory dimensions to 50, and used AUC [11] as our evaluation metric.

4.2 Performance Prediction

This experiment addresses RQ1 by comparing THMN with seven state-of-the-art KT models across four educational datasets in terms of AUC. Table 1 shows that THMN consistently outperforms all baseline models in all data sets.

Performance improvements are particularly notable on Kddcup2010 and Synthetic-5 datasets, which contain complex concept hierarchies and interdependencies. This suggests that THMN's approach is especially valuable for educational domains with intricate prerequisite structures and multi-concept questions.

Table 1. AUC comparison of THMN with seven baseline and advanced KT models.

Model	Dataset			
	Statics2011	Kddcup2010	Synthetic-5	ASSIST2009
DKT	80.20	78.63	80.36	80.53
DKVMN	82.84	79.81	82.74	81.63
GKT	73.43	76.92	74.27	72.31
SKT	74.21	78.16	76.31	74.63
DGEKT	79.54	77.2	79.95	76.68
HawkesKT	80.8	79.51	82.25	76.25
KTMFF	82.97	80.11	83.25	81.76
THMN	**83.42**	**80.63**	**83.74**	**82.17**

[1] https://sites.google.com/site/assistmentsdata/datasets/assistments-problems.

4.3 Evolution of Knowledge States

To answer RQ2, we visualize the knowledge states of concepts captured by DKVMN, the primary baseline, and THMN over a sequence of 50 exercises from the Synthetic-5 dataset in Fig. 2. A key difference lies in how each model responds to incorrect answers. For instance, at time steps $t = 15$ and $t = 36$, THMN shows a realistic, gradual decline in the knowledge state of concept c_2, reflecting natural learning dynamics. In contrast, DKVMN increases the knowledge state despite errors that contradict fundamental educational principles of assessment. THMN also incorporates practice-sensitive knowledge updates, unlike DKVMN. At $t = 34$, after an incorrect response, THMN applies only a slight reduction, accounting for prior engagement with the concept, consistent with cognitive science research on knowledge stability. This highlights how THMN's adaptive scaling mechanism fosters more robust knowledge representations by considering practice diversity, a key principle in learning theory.

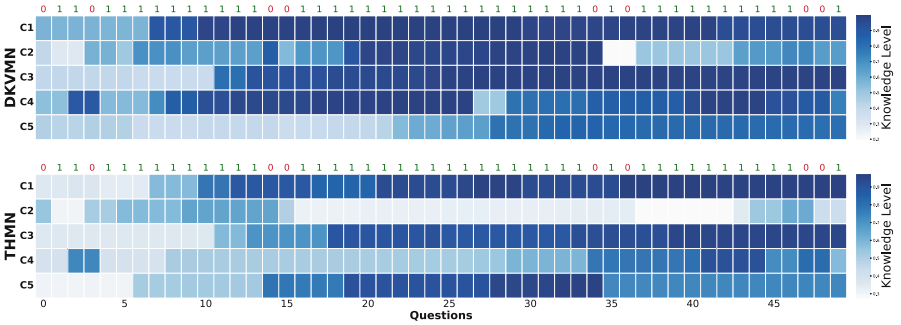


Fig. 2. Knowledge evolution of a student on five concepts during a 50-question sequence from Synthetic-5, captured by (a) DKVMN and (b) THMN models. Rows represent the knowledge evolution of concepts while columns show sequential responses to questions.

5 Conclusion and Future Work

We proposed THMN, enhancing knowledge tracing through temporal hypergraphs that capture dynamic concept interactions. THMN improves prediction accuracy by modeling higher-order relationships via bidirectional message passing and integrating heterogeneous information through attention fusion, while enhancing knowledge modeling through an adaptive scaling mechanism based on practice diversity. Experiments across four datasets show THMN's superiority over state-of-the-art models in both prediction accuracy and knowledge evolution realism. Future work will address scalability and inference optimization.

References

1. Abdelrahman, G., Wang, Q.: Deep graph memory networks for forgetting-robust knowledge tracing. IEEE Trans. Knowl. Data Eng. **35**(8), 7844–7855 (2022)
2. Abdelrahman, G., Wang, Q., Nunes, B.: Knowledge tracing: a survey. ACM Comput. Surv. **55**(11), 1–37 (2023)
3. Abdi, S., Khosravi, H., Sadiq, S., Gasevic, D.: A multivariate Elo-based learner model for adaptive educational systems. arXiv preprint arXiv:1910.12581 (2019)
4. Antelmi, A., Cordasco, G., Polato, M., Scarano, V., Spagnuolo, C., Yang, D.: A survey on hypergraph representation learning. ACM Comput. Surv. **56**(1), 1–38 (2023)
5. Cencetti, G., Battiston, F., Lepri, B., Karsai, M.: Temporal properties of higher-order interactions in social networks. Sci. Rep. **11**(1), 7028 (2021)
6. Choi, Y., et al.: Towards an appropriate query, key, and value computation for knowledge tracing. In: Proceedings of the Seventh ACM Conference on Learning@ Scale, pp. 341–344 (2020)
7. Cui, C., et al.: DGEKT: a dual graph ensemble learning method for knowledge tracing. ACM Trans. Inf. Syst. **42**(3), 1–24 (2024)
8. Dong, Y., Sawin, W., Bengio, Y.: HNHN: Hypergraph networks with hyperedge neurons. arXiv preprint arXiv:2006.12278 (2020)
9. 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)
10. Käser, T., Klingler, S., Schwing, A.G., Gross, M.: Dynamic bayesian networks for student modeling. IEEE Trans. Learn. Technol. **10**(4), 450–462 (2017)
11. Ling, C.X., Huang, J., Zhang, H., et al.: AUC: a statistically consistent and more discriminating measure than accuracy. In: IJCAI. vol. 3, pp. 519–524 (2003)
12. Minn, S., Yu, Y., Desmarais, M.C., Zhu, F., Vie, J.J.: Deep knowledge tracing and dynamic student classification for knowledge tracing. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 1182–1187. IEEE (2018)
13. Nakagawa, H., Iwasawa, Y., Matsuo, Y.: Graph-based knowledge tracing: modeling student proficiency using graph neural network. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 156–163 (2019)
14. Piech, C., et al.: Deep knowledge tracing. Adv. Neural Inf. Process. Syst. **28** (2015)
15. Shen, S., et al.: A survey of knowledge tracing: Models, variants, and applications. IEEE Transactions on Learning Technologies (2024)
16. Song, X., Li, J., Cai, T., Yang, S., Yang, T., Liu, C.: A survey on deep learning based knowledge tracing. Knowl.-Based Syst. **258**, 110036 (2022)
17. Tong, S., et al.: Structure-based knowledge tracing: an influence propagation view. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 541–550. IEEE (2020)
18. Wang, C., et al.: Temporal cross-effects in knowledge tracing. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 517–525 (2021)
19. Xiao, Y., Xiao, R., Huang, N., Hu, Y., Li, H., Sun, B.: Knowledge tracing based on multi-feature fusion. Neural Comput. Appl. **35**(2), 1819–1833 (2023)
20. Yang, Y., et al.: GIKT: a graph-based interaction model for knowledge tracing. In: Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, proceedings, part I, pp. 299–315. Springer (2021)

21. 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
22. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th International Conference on World Wide Web, pp. 765–774 (2017)



An Augmented Intelligence System for Automated Quality Control and Feedback Generation of Multiple Choice Test Items

Tahereh Firoozi(✉), Lia Daniels, Vijay Daniels, and Mark Gierl

University of Alberta, Edmonton, AB T6G2R3, Canada
{tahereh.firoozi,lia.daniels,vijay.daniels,mark.gierl}@ualberta.ca

Abstract. Writing high-quality multiple-choice (MC) test items is crucial for accurately assessing student's knowledge and skills. Linguistic and formatting considerations are critical for writing high-quality items, yet they are often overlooked by item writers. Manually verifying the linguistic and formatting guidelines for every MC item is both time-consuming and prone to error. This study introduces an augmented intelligence multi-agent system that detects item-writing flaws, provides automated feedback, and offers AI-generated suggestions to help item writers produce high-quality MC items. The system leverages established item-writing guidelines for flaw detection and incorporates GPT-4 to generate feedback and recommend corrections. Results indicate the system was accurate on three criteria: 85% accuracy in producing flawless items, 90% accuracy in ensuring linguistic clarity, and 70% accuracy in maintaining content validity. These findings support the use of augmented intelligence systems for creating high-quality MC test items.

Keywords: Item writing quality · Multiple-choice test item · Augmented intelligence system

1 Introduction

Multiple-choice (MC) items are widely used across educational settings because they are efficient to administer and reliable to score. Ensuring the quality of MC test items is essential for accurately measuring students' true abilities [1]. Haladyna and Downing's list of item-writing guidelines [2], has shaped scholarly discourse on how to structure and write high-quality MC items. Despite the prominence of these guidelines, adherence by item-writers remains inconsistent. Numerous studies have shown that teacher-made tests, at various educational levels, often include item-writing flaws that deviate from Haladyna and Downing's recommendations [3,4]. Many of these flaws pertain to the language and format of MC items. Research indicates that item-writers frequently overlook linguistic and formatting considerations when developing MC items [3]. Manually verifying each guideline for every MC item is time-consuming and prone

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 86–93, 2025.
https://doi.org/10.1007/978-3-031-99267-4_11

to human error, particularly with large item banks [1]. Moreover, widely used automated metrics for MC item quality, such as BLEU and ROUGE, do not fully address the linguistic and educational value of the items [5,6]. Recently, researchers have explored both rule-based and large language model (LLM)-based systems [6,7] to evaluate the quality of MC test items according to item-writing flaw rubrics. One such evaluation toolkit [7] significantly advanced educational testing by accurately detecting linguistic and educational flaws in MC test items. However, its authors noted that future work could focus on using LLMs to provide detailed feedback on how to correct identified flaws. In this study, we address this gap by introducing an augmented intelligence multi-agent system that can detect item-writing flaws, provide automated feedback, and offer AI-generated item suggestions to assist item-writers in creating high-quality items. The proposed framework leverages an AI multi-agent system that employs established item-writing guidelines to detect flaws, and GPT-4's capabilities to generate feedback and corrections. By combining rule-based AI with a generative LLM, the system can make informed, consistent decisions when providing feedback and suggestions for flawed MC items.

2 Background

2.1 Related Works

With the advent of LLMs, research has increasingly focused on both the challenges and the potential of these generative models for automatic MC item generation [8,9]. Although studies showed that LLMs are reliable for generating high-quality items, they often perform poorly in identifying the correct answers for the generated items [7,8]. Research further indicates that the quality of the AI generated MC items varies across content areas. For instance, a study on computing science courses revealed that GPT-3.5 can automatically generate items aligned with the course outline [10]. However, another study [9] found that GPT-3.5 failed to produce quality MC items in more than half of the generated items within the domain of immunology.

Given the interest in evaluating the performance of the AI generated items, recent work [5–7] underscores the importance of human-based guidelines for evaluating item quality. These researchers argue that common metrics, such as BLEU and ROUGE, do not adequately capture the linguistic quality of items. For example, Grvisse [5] used three different GPT-based systems to generate 150 MC items, then assessed them against human-based guidelines. Results showed that 37% of the items contained at least one flaw, with the most common issues including ambiguous keys and "all of the above" options. Moore et al. [7], similarly emphasized the limitations of purely quantitative metrics for MC item evaluation and highlighted the importance of using human evaluation as a benchmark for accuracy. They developed a human-centered MC item evaluation toolkit based on the rules in item-writing guidelines. These rules were employed into the system using NLP techniques and GPT-4 model and tested on 100 MC items. When comparing evaluations by humans versus the system, the system

was generally more robust, yet agreement between humans and the system was high ($\kappa = 75.3$). Building on Moore et al. [7], in this study we introduce a multi-agent system that first detects item flaws and then generates suggested items to revise flawed items using GPT-4 API.

2.2 Item Writing Guideline

In a seminal article, Haladyna et al. [11] used empirical evidence and expert consensus to create the most cited item writing guidelines in the literature of educational testing. In this study, we focus on six item writing flaws that specifically address the language and formatting of MC item stems and options [12]:

1. Format the options vertically and order them by size.
2. Phrase the stem positively and avoid using negatives (e.g., not/except).
3. If negatives are necessary, write them in boldface or capitalized.
4. Limit the number of choices between three to five.
5. Avoid "All of the above," "None of the above," and combinations of options.
6. Use standard stem formats, such as complete sentence, question or best-answer versions.

For this study, we defined specific rules to operationalize each of these six guidelines, and integrated them into the architecture of our AI agent system.

3 System Overview

The system is designed to first identify and flag any violations of the six defined rules in human-written MC test items. It then generates a revised version of the original item, including the correct key option. The workflow (Fig. 1) begins with a detector agent that evaluates the quality of each MC item based on the six defined rules. These rules are implemented as tools which the agent invokes asynchronously during its decision-making process to determine whether an item "passes" or "fails." When no flaws are detected, the detector agent sends the item through the "pass" route, wherein the original item and its key option are output by the system. However, if the item contains any flaws, it is routed to the "fail' path, leading to the LLM boxwhich serves as the system's language enginefor revision. In our framework, the LLM box is powered by GPT-4o and is guided by the defined rules and the context of the original item.

To enhance the accuracy of the GPT-4o model, we employed a chaining strategy [13] to break down the three tasks in the LLM box into sequential steps. With this approach, the output from one step serves as the input for the next. In the first chain, the model receives two inputs: the original item and the defined rules. The model is prompted to produce (1) a summary of any flaws in the item based on these rules, and (2) a revised, flaw-free version of the item while minimizing changes to the original item. In the second chain, the GPT-4o model is provided with three inputs: the original item, the original key option, and the item generated in the previous chain. This step determines the correct

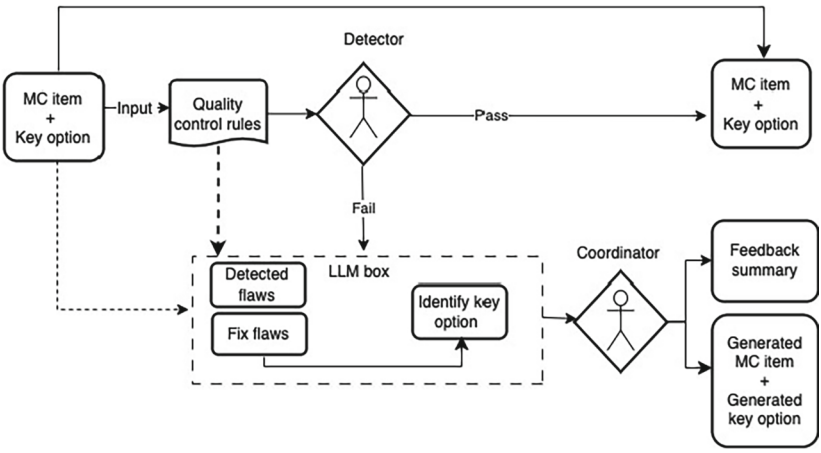


Fig. 1. The system workflow.

key option for the revised item, ensuring that rewording or reordering does not inadvertently alter or remove the original correct answer. Finally, a coordinator agent consolidates the outputs from these chain steps and the rule functions. The result is the system's final output, which includes a summary of the detected flaws and a generated MC item with its corresponding keyed option.

4 Experiment

4.1 Data

In this pilot study, we randomly selected 20 multiple-choice items from an item bank of 180 items, which had originally been created by a university instructor to cover eight chapters in an introductory undergraduate Macroeconomics course. For each testing outcome item in the bank, the instructor developed an average of three different items, from which we randomly selected one item.

4.2 Human Evaluation

A panel of human experts evaluated the system at two levels: (1) item quality of the generated feedback item, and (2) content validity of the generated items. For the item quality evaluation, two experts with extensive experience as instructors and scholars in item writing performed the following tasks: (1) They assessed whether the system-generated items were free from flaws, according to the item-writing guidelines used in this study, and (2) They reviewed the grammatical and linguistic quality of each generated item, using a four-point scale (1. accept, 2. accept with minor revisions, 3. reject with major revisions, and 4. reject) in [14] for evaluating the generated items.

The interrater reliability for the first task was high ($\kappa = 0.81$). The only disagreement between the raters arose in the "Option ordering flaw" category (Fig. 2). Rater 1 identified three generated items in which the options were not ordered; whereas, Rater 2 flagged five such items. Subsequently, one of the authors in this studyan expert in educational testingexamined the items labeled as having an "option ordering flaw." This review revealed that all options in the generated items were arranged in ascending order based on their character count, as defined by the system. Therefore, the system correctly arranged the options based on the defined rules.

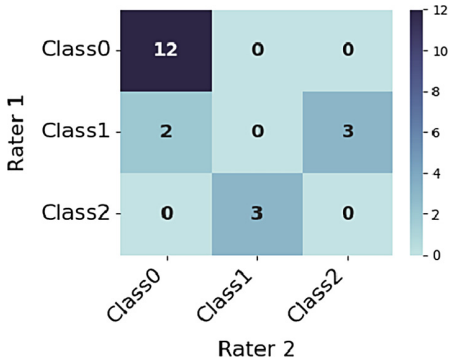


Fig. 2. Confusion matrix for the rating of the generated items. *Note: Class0 = flawless items, Class1 = items with option ordering flaw, Class2 = items with disallowed option flaw.*

In the second task, there were frequent discrepancies between the raters for adjacent categoriesspecifically between scales 1 and 2, and between scales 3 and 4. To address these discrepancies, we applied a Quadratic Weighted Kappa (QWK) analysis, resulting in an agreement of 82%. The raters discussed their disagreements, and the system performance was evaluated only after they reached a consensus on the disputed items. For content-level evaluation, a subject matter expert (SME) reviewed the system generated items and determined whether each item aligned with the intended content of the original item, using the guiding question, "Does the new item still measure the content you intended?". The SME spontaneously documented the reasons for any items that were considered misaligned or failed to meet the content criteria.

5 Results

5.1 Generating Flawless Feedback Items

The system successfully detected and corrected item-writing flaws in 17 out of the 20 items (85%). The three items that the system generated poorly contained disallowed options (i.e., "All of the above," "None of the above," or combined options such as "A and B"). In addition, the system correctly identified the key options for all of the generated items.

5.2 Content Validity

The content rating by the SME revealed that 14 out of 20 generated items measured the same content that the instructor intended to measure. One of the items that did not meet the content validity criterion had a problematic option; as the SME noted "The wording for the answer [options] seems more confusing than the original ones." Three of the failed generated items had unacceptable stems and two of the items had both unacceptable stem and options.

6 Discussion and Conclusion

Results of this study indicate that the system performed accurately on more than 70% of the items on the three evaluation criteria, including the generation of flawless items, ensuring linguistic clarity, and maintaining content validity. Consistent with previous research [10], our results reveal that when LLMs are augmented with human-defined rules, they can yield more reliable test items compared to using LLMs alone as the primary knowledge source [9]. Despite these positive outcomes, a key weakness of the system is in handling disallowed options when generating feedback items to fix this flaw. The system relies on identifying conventional patterns, such as "none of the above" or "both A and B." However, some original items in this study contained disallowed options that did not follow these patterns by merging two or more concepts without explicitly using a recognizable disallowed-option format. One explanation for this shortcoming is the constraint we placed on GPT-4o, which strictly adheres to predefined rules and the original item's content. While this approach increases reliabilityby ensuring alignment with the instructor's intended contentit reduces the model's flexibility [15]. Given the promising results from multi-stage prompting (MSP) approaches for MC item generation [16], future work could explore integrating MSP into our augmented intelligence system. This integration may enhance flexibility without compromising the reliability demonstrated in the current study.

Human evaluation indicated that all but two of the generated items were considered linguistically acceptable. Both problematic items contained confusing stems and some redundant elements in the options. The same items were also flagged by the rater for having a confusing stem. Similar to the disallowed option flaw, this redundancy may be attributed to the system's constraints to ensure that the LLM adheres to the content of the original items. As a possible solution, expanding the item-writing guidelines within the systemparticularly by adding rules against overlapping words in the options and stemscould improve performance [7]. The confusing stem may also reflect the system's limited domain knowledge. Given that the original items serve as the primary source of content knowledge, integrating lecture notes or textbook materials could further enhance LLM output [17]. Contrary to prior research [8], the GPT-4o used in our system consistently provided accurate key options for the generated items. Notably, restricting the system to the original item's content may have contributed to its high accuracy in identifying key options. This outcome can also underscore the

value of a multi-step (chaining) approach [13], which enhances LLM reliability by breaking the task into multiple steps.

It is important to note that the set of rules employed by the system is limited to the language and structure of MC items. We plan to expand the system's scope by incorporating additional guidelines from Haladyna et al. [11]. Moreover, future work will evaluate the system's performance adding quantitative metrics to the current results. We also aim to compare results across different volumes of MC items and a broader range of content areas.

References

1. de Fitero-Dominguez, D., Garcia-Cabot, A., Garcia-Lopez, E.: Automated multiple-choice question generation in Spanish using neural language models. In: Neural Computing and Applications, vol. 36, no. 29, pp. 18223–18235 (2024)
2. Haladyna, T.M., Downing, S.M: Validity of a taxonomy of multiple choice item-writing rules. In: Applied Measurement in Education, vol. 2, no. 1, pp. 51–78 (1989)
3. Tarrant, M., Ware, J.: Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. In: Medical Education, vol. 42, no. 2, pp. 198–206 (2008)
4. Tarrant, M., et al.: The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. In: Nurse Education Today, vol. 26, no. 8, pp. 662–671 (2006)
5. Grévisse, C.: Comparative quality analysis of GPT-based multiple choice question generation. In: International Conference on Applied Informatics. Springer, pp. 435–447 (2023)
6. Moore, S., et al.: Assessing the quality of multiple-choice questions using GPT- 4 and rule-based methods. In: European Conference on Technology Enhanced Learning. Springer, pp. 229–245 (2023)
7. Moore, S., et al.: An automatic question usability evaluation toolkit. In: International Conference on Artificial Intelligence in Education. Springer, pp. 31–46 (2024)
8. Tran, A., et al.: Generating multiple choice questions for computing courses using large language models. In: IEEE Frontiers in Education Conference (FIE), pp. 1–8. IEEE (2023)
9. Ngo, A., et al.: ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. In: Academic Pathology, vol. 11, no. 1, p. 100099 (2024)
10. Song, T., et al.: Automatic generation of multiple-choice questions for CS0 and CS1 curricula using large language models. In: International Conference on Computer Science and Education. Springer, pp. 314–324 (2023)
11. Haladyna, T.M., Downing, S.M., Rodriguez, M.C.: A review of multiple-choice item-writing guidelines for classroom assessment. In: Applied Measurement in Education, vol. 15, no. 3, pp. 309–333 (2002)
12. Daniels, L., et al.: Multiple-choice item-writing guidelines for classroom assessment: a state-of-the-art review and use-inspired reframing. In: Educational Assessment, Evaluation and Accountability (in press)
13. Wu, T., Terry, M., Cai, C.J.: AI chains: transparent and controllable human-AI interaction by chaining large language model prompts. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–22 (2022)

14. Gierl, M.J., Lai, H., Tanygin, V.: Methods for validating generated items: a focus on model-level outcomes. In: Advanced Methods in Automatic Item Generation, pp. 120–143 (2021)
15. Babu, M.V.S., Banana, K.: A study on narrow artificial intelligence—an overview. Int. J. Eng. Sci. Adv. Technol. **24**(4), 210–219 (2024)
16. Maity, S., Deroy, A., Sarkar, S.: A novel multi-stage prompting approach for language agnostic MCQ generation using GPT. In: European Conference on Information Retrieval. Springer, pp. 268–277 (2024)
17. Darshan, P., et al.: Leveraging LLM and RAG for automated answer script evaluation. In: 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), pp. 1–5 (2024)

ClickSight: Interpreting Student Clickstreams to Reveal Insights on Learning Strategies via LLMs

Bahar Radmehr[1(✉)], Ekaterina Shved[1], Fatma Betül Güreş[1], Adish Singla[2], and Tanja Käser[1]

[1] EPFL, Lausanne, Switzerland
{bahar.radmehr,kate.kutsenok,fatma-betul.gures,tanja.kaeser}@epfl.ch
[2] MPI-SWS, Saarbrücken, Germany
adishs@mpi-sws.org

Abstract. Clickstream data from digital learning environments provide valuable insights into student behavior but are challenging to interpret due to their granularity. Prior methods mainly relied on handcrafted features, expert labeling, clustering, or supervised models, limiting generalizability and scalability. We present ClickSight, an in-context Large Language Model (LLM)-based pipeline that interprets student clickstreams given a list of learning strategies to generate textual interpretations of students' behaviors during interaction. We evaluate four prompting strategies and assess the effect of self-refinement across two open-ended environments using domain-expert rubric-based evaluations. Results show that while LLMs can reasonably interpret learning strategies from clickstreams, interpretation quality varies by prompting strategy, and self-refinement offers limited improvement. ClickSight demonstrates the potential of LLMs to generate theory-driven insights from educational interaction data.

Keywords: learning strategies · student clickstreams · large language models · log data · open-ended learning environments

1 Introduction

Digital learning environments have enabled more personalized, scalable, and data-driven education [10]. They facilitate gathering rich clickstream data—detailed, time-stamped records of student interactions—which can offer valuable insights into learning behaviors. Such insights support researchers in further understanding student behavior, help instructors provide timely interventions, and guide students in improving their learning strategies [3]. Yet, interpreting clickstreams remains challenging due to their high dimensionality and granularity that makes expert annotation labor-intensive and cognitively demanding [6].

Prior work in AI for education has employed various methods to detect learning strategies and analyze behavior from clickstreams, including domain-specific

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 94–102, 2025.
https://doi.org/10.1007/978-3-031-99267-4_12

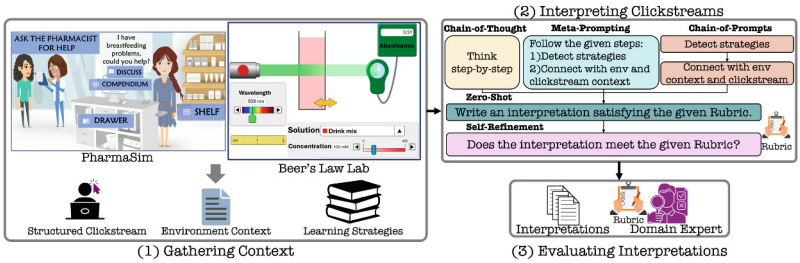


Fig. 1. The ClickSight pipeline interprets student clickstreams in three stages. (1) clickstream data from PharmaSim and Beer's Law Lab are structured, and context and learning strategies are gathered; (2) an LLM is prompted using one of four prompting strategies (*Zero-shot*, *Chain-of-Thought*, *Meta-Prompting*, *Chain-of-Prompts*) with optional self-refinement to generate interpretations; (3) human experts assess the outputs using a quality rubric.

feature design [7,19], behavioral patterns-based clustering [19], sequential pattern mining to uncover frequent interaction traces [1,21], and supervised classification [1]. Though insightful, these approaches often demand extensive manual work and domain expertise, and poorly generalize across contexts.

LLMs like GPT-4o [11] offer new opportunities in educational technology [4], with applications including content generation [17], feedback provision [12], and student simulation [13]. Given their strong ability to recognize patterns [9], LLMs hold promise for interpreting clickstream data through the lens of learning strategies. However, this potential remains largely unexplored.

We introduce ClickSight, an in-context LLM-based pipeline that interprets student clickstreams using predefined learning strategies. Given a clickstream and learning strategy list, it generates behavioral interpretations. We consider four prompting strategies and evaluate self-refinement, where the LLM critiques and revises its own output. ClickSight is evaluated in two domains: **PharmaSim** [14], a pharmacy assistant diagnostic training simulation, and **Beer's Law Lab** [19], a virtual chemistry lab for secondary students. To evaluate interpretation quality, we create a rubric grounded in human-centered explanation theory and conduct domain expert evaluations. Our research questions are: (1) Which prompting strategy yields the best strategy-based interpretation **(RQ1)**? (2) How does self-refinement affect interpretation quality **(RQ2)**?

Our results show that ClickSight enables high-quality interpretations, with one prompting strategy outperforming the rest. Self-refinement yields modest gains for some strategies but introduces errors in others. Our implementation is available at https://github.com/epfl-ml4ed/ClickSightLBR25.

2 Methodology

ClickSight consists of three stages (see Fig. 1). In the first stage, we describe the simulation setting, identify relevant learning strategies, and organize student

clickstreams into a structured format. The next stage prompts an LLM to generate textual interpretations grounded in the contextual information constructed in the first stage. Finally, we apply a rubric-based domain-expert evaluation to assess the quality, clarity, and relevance of the generated interpretations.

2.1 Gathering Environment Context

We consider two different environments PharmaSim [14] and Beer's Law Lab [19].
PharmaSim is a scenario-based learning environment for pharmacy assistant training, where students build diagnostic skills through simulated interactions with a virtual patient. In the chosen scenario (Fig. 1), a mother visits the pharmacy with a breastfeeding issue. The student investigates by asking questions, consulting the pharmacist, researching medicines, examining shelf items, and reviewing medical documentation.
Student Clickstreams for PharmaSim. Following prior work [14], we represent each clickstream as a sequence of function calls tied to action types in the simulation. For instance, asking the virtual mother about her symptoms at time t is encoded as: `discuss(mother, symptoms,` t`) [output: My breast hurts...]`.
Learning Strategies for PharmaSim. We identified **nine** diagnostic strategies from prior work, reflecting behaviors observed in environments like PharmaSim. These strategies vary in effectiveness covering, different environment aspects. Effective conversation strategies [5] include *LINDAAFF*, where students ask focused symptom-related questions, and *Inquiry about Relevant External Factors*, which explores external influences like the baby's health. Ineffective approaches include *Premature Closure* (settling on a diagnosis too early), *Random Inquiry* (asking irrelevant questions), and *Insufficient Inquiry* (incomplete information gathering). *Research strategy* refers to how students use documentation, compendium, and shelf. Drawing on self-regulation theory [22], we define three types: Targeted Research (focused use), Unfocused Research (scattered browsing), and Minimal Research (little to no use). *Hint-seeking strategy* includes Thoughtful Hint Seeking (after meaningful effort), Premature Hint Seeking (too early without sufficient exploration), and No Hint Seeking (no requests) [18]. Students may also engage in *Iterative Reflection*, pausing to process information [15], or *Gaming the System*, rushing through tasks without meaningful engagement [2].
Beer's Law Lab lets students explore how cuvette width, solution concentration, and laser color-solution combinations affect light absorbance (Fig. 1). Width and concentration have a linear relationship with absorbance; laser-solution color shows a non-linear pattern.
Student Clickstreams for Beer's Law Lab. Students perform two action types: *exploration*, manipulating variables and observing absorbance changes, and *analysis*, recording values. We focus on exploration actions for strategy-based interpretations, formatted as function calls. For example: `explore(variable=width, change=`$c \rightarrow d$`, start=`t`, duration=`s`) [absorbance:` $c \rightarrow d$`]` , with consecutive actions on the same variable within 3 s are merged [19].

Learning Strategies for Beer's Law Lab. We analyzed student behavior using three core inquiry strategies from prior work [19]: *Control of Variables Strategy (CVS)*, which is isolating one variable by holding others constant; *Range*, which is exploring the full value span to reveal trends; and *Optimal*, which is selecting non-focal values to minimize confounds (e.g., avoiding laser-solution color match when testing cuvette width). These strategies vary across variables and over time, making them harder to interpret than the aggregated strategies in PharmaSim.

2.2 Interpreting Clickstreams

At the core of ClickSight is an LLM that interprets student clickstreams using the strategies and context from stage 1, guided by a rubric grounded in human-centered explanation research. We explored four prompting strategies–*Zero-shot*, *Chain-of-Thought*, *Meta-Prompting*, and *Chain-of-Prompts*, and added a self-refinement step where the LLM reviews and improves its responses.

Table 1. Quality evaluation rubric for clickstream interpretation.

Criterion	Definition
Completeness	Percentage execution level of learning strategies across different aspects mentioned in the interpretation.
Correctness	Percentage of correct execution levels of strategies stated in the interpretation based on the clickstream.
Justifiedness	Percentage of conclusions about strategy execution levels linked to evidence in the clickstream.
Comprehensibility	Whether the explanation is clear and easy to understand, without ambiguity.

Rubric for High-Quality Interpretations. To guide the LLM in generating meaningful interpretations, we included a detailed description of what constitutes a high-quality interpretation directly in the prompt. Based on prior work in human-centered explanation [20], the rubric specifies four criteria (Table 1): *Completeness*, *Correctness*, *Justifiedness*, and *Comprehensibility*. *Completeness* requires full coverage of strategy execution levels across environment aspects. *Correctness* ensures alignment between stated execution levels and clickstream evidence, and penalizes missing mentions. *Justifiedness* demands that conclusions be supported by clickstream evidence (e.g., actions). *Comprehensibility* assesses clarity and is unmet if the interpretation is ambiguous. This rubric was used to shape LLM responses and assess quality of the generated interpretations.

Prompting Strategies. To study how prompting affects interpretation quality, we designed four strategies, *Zero-shot*, *Chain-of-Thought*, *Meta-Prompting*, and *Chain-of-Prompts*, based on prior work in effective LLM prompting [16]. Each prompt includes the student's clickstream, environment context, learning strategies, and the rubric. Detailed prompts are available on our GitHub.

- *Zero-shot*: Instructs LLM to produce an interpretation that satisfies the rubric, without reasoning guidance, relying on its internal task knowledge.
- *Chain-of-Thought*: Adds an instruction to encourage incremental reasoning.
- *Meta-Prompting*: Gives multi-step instructions in a single prompt to detect strategy execution levels, link to context, and generate an interpretation.
- *Chain-of-Prompts*: Breaks the task into three prompts to detect strategy execution levels, link to context, then generate an interpretation.

Self-refinement. To better align outputs with rubric criteria, we add a self-refinement step applicable to any prompting strategy. After generating an initial interpretation, the LLM evaluates it using binary questions aligned with the four rubric dimensions. *Completeness*, *Correctness*, and *Justifiedness* are each assessed with multiple questions (one per strategyaspect pair); *Comprehensibility* uses a single question. If any response is negative, the LLM produces targeted feedback and revises its output. This process can repeat; we use three rounds, which sufficed for meaningful improvement. We refer to the initial version as *Initial* and the final version as *Self-Refined*.

2.3 Evaluating Interpretations

In this stage, human experts evaluate the generated interpretations using the rubric criteria. Two independent learning science experts per environment rated each output with a structured grading scheme reflecting how well it satisfies each rubric dimension.

Grading Scheme. Following the approach in 2.2, we used binary questions to guide expert evaluation of each rubric criterion. *Completeness*, *Correctness*, and *Justifiedness* were each assessed with nine targeted questions, one per strategy-aspect pair: PharmaSim has 9 strategies in one aspect; Beer's Law Lab has 3 strategies across 3 aspects. *Comprehensibility* was evaluated with one question.

Grading Process. To evaluate each behavior profile identified through clustering, we sampled five students per cluster, resulting in 30 instances for PharmaSim (6 clusters from k-means, k via elbow method) and 20 for Beer's Law Lab (4 clusters, following [19]). This produced 400 interpretations (50 students $\times$ 4 prompting strategies $\times$ 2 self-refinement conditions) for annotation. All participants gave informed consent, and the study was approved by the ethics committee (Nr. 010-2023, 062-2021). To ensure consistency, two experts independently evaluated one representative clickstream per cluster across all prompting strategies, totaling 24 samples for PharmaSim and 16 for Beer's Law Lab. We measured inter-annotator agreement using Cohen's κ for each binary question [8], reporting both average and minimum scores for multi-question criteria. For PharmaSim, the agreement scores were: $\kappa_{\text{avg-complete}} = 1.0$, $\kappa_{\text{avg-correct}} = 0.95$ ($\kappa_{min} = 0.74$), $\kappa_{\text{avg-justified}} = 0.97$ ($\kappa_{min} = 0.86$), and $\kappa_{\text{comprehensible}} = 1.0$, and for Beer's

Law Lab: $\kappa_{\text{avg-complete}} = 1.0$ ($\kappa_{min} = 1.0$), $\kappa_{\text{avg-correct}} = 0.98$ ($\kappa_{min} = 0.89$), $\kappa_{\text{avg-justified}} = 0.91$ ($\kappa_{min} = 0.70$), and $\kappa_{\text{comprehensible}} = 1.0$. Following high agreement, remaining annotations were completed by one expert.

Score Calculation. Criterion-level scores were computed by averaging binary responses for multi-question criteria. *Comprehensibility*, evaluated with a single yes/no question, was scored directly (1 for clear, 0 for ambiguous interpretations). To compute overall quality (01), we first collected binary responses to nine questions across the three multi-question criteria (*Completeness*, *Correctness*, *Justifiedness*). Each strategy received a composite score by multiplying its three criterion scores, rewarding correct and justified mentions. We averaged the nine strategy scores and multiplied the result by the *Comprehensibility* score, ensuring overall quality required clarity.

3 Results

We evaluated ClickSight on PharmaSim and Beer's Law Lab clickstreams to assess how prompting strategy (RQ1) and self-refinement (RQ2) affect interpretation quality.

3.1 Comparative Evaluation of Prompting Strategies (RQ1)

We compared interpretation quality across the four prompting strategies, *Zero-shot*, *Chain-of-Thought*, *Meta-Prompting*, and *Chain-of-Prompts*, using our evaluation rubric (Sect. 2.3). Table 2 shows average scores (with standard deviations) across all criteria and environments. *Zero-shot* achieved the highest overall score in both environments and a perfect *Completeness* score, while *Chain-of-Thought* ranked lowest on this criterion due to omissions of some strategies in its step-by-step reasoning. In *Correctness*, *Meta-Prompting* performed best in PharmaSim, likely due to processing all strategies in one prompt, but this did not transfer to Beer's Law Lab. *Chain-of-Thought* and *Meta-Prompting* also scored lower in *Justifiedness*, often producing shorter outputs with insufficient clickstream evidence. All strategies achieved perfect *Comprehensibility* in PharmaSim, but only *Zero-shot* did so in Beer's Law Lab. Other outputs occasionally included raw, function-style elements that hindered readability, despite all prompts instructing the LLM to describe student actions in natural language.

Table 2. Performance of different prompting strategies on PharmaSim and Beer's Law Lab.

	PharmaSim					Beer's Law Lab				
Prompting Strategy	*Completeness*	*Correctness*	*Justifiedness*	*Comprehensibility*	**Overall**	*Completeness*	*Correctness*	*Justifiedness*	*Comprehensibility*	**Overall**
Zero-shot	1.00 ±0.00	0.79 ±0.02	1.00 ±0.00	1.00 ±0.00	**0.79 ±0.02**	1.00 ±0.00	0.76 ±0.04	1.00 ±0.00	1.00 ±0.00	**0.76 ±0.04**
Chain-of-Thought	0.93 ±0.02	0.80 ±0.03	0.87 ±0.03	1.00 ±0.00	0.66 ±0.04	0.87 ±0.05	0.67 ±0.05	0.19 ±0.07	0.85 ±0.08	0.15 ±0.07
Meta-Prompting	0.96 ±0.01	0.89 ±0.02	0.74 ±0.04	1.00 ±0.00	0.61 ±0.03	0.99 ±0.01	0.72 ±0.04	0.83 ±0.06	0.90 ±0.07	0.56 ±0.06
Chain-of-Prompts	0.96 ±0.02	0.77 ±0.02	1.00 ±0.00	1.00 ±0.00	0.74 ±0.03	0.99 ±0.01	0.72 ±0.04	0.98 ±0.01	0.95 ±0.05	0.68 ±0.05

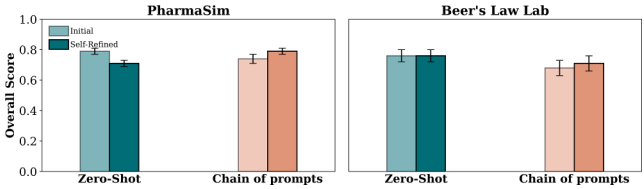


Fig. 2. Overall scores of *Initial* and *Self-Refined* interpretations generated using *Zero-shot* and *Chain-of-Prompts* prompting strategies in PharmaSim and Beer's Law Lab.

3.2 Effect of Self-refinement (RQ2)

In a second analysis, we examined the impact of self-refinement on interpretation quality. We applied self-refinement to the outputs from the two best-performing strategies, *Zero-shot* and *Chain-of-Prompts*. Figure 2 shows the overall score for both *Initial* and *Self-Refined* interpretations. Refinement improved *Chain-of-Prompts*, especially in *Completeness* and *Correctness*, across both environments. However, it slightly reduced *Zero-shot* performance in PharmaSim due to hallucinations during rubric self-evaluation. In Beer's Law Lab, it had no effect because the LLM assumed the initial output already satisfied all criteria.

4 Conclusion

In this paper, we introduced ClickSight, an in-context LLM framework for interpretation of student clickstreams, along with a rubric grounded in human-centered explanation theory. We evaluated four prompting strategies, with and without self-refinement, across two educational environments. ClickSight enabled LLMs to generate high-quality interpretations, with *Zero-shot* performing best. Self-refinement showed mixed results, pointing to the need for refinement methods tailored to the task. While our study focused on zero-shot prompting, future work should explore few-shot settings and adapt outputs for stakeholders such as teachers and students. Overall, ClickSight demonstrates the potential of LLMs to generate theory-aligned insights from interaction data.

Acknowledgements. This project was substantially funded by the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

1. Akpinar, N., Ramdas, A., Acar, U.A.: Analyzing student strategies in blended courses using clickstream data. In: Rafferty, A.N., Whitehill, J., Romero, C., Cavalli-Sforza, V. (eds.) Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020. International Educational Data Mining Society (2020). https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_172.pdf

2. Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: Gaming the system: learning from mixed-response data. In: Proceedings of the 8th International Conference on Intelligent Tutoring Systems, pp. 272–281. Springer (2006)
3. Baker, R.S., Inventado, P.S.: Educational Data Mining and Learning Analytics, pp. 61–75. Springer New York, New York, NY (2014). https://doi.org/10.1007/978-1-4614-3305-7_4
4. Denny, P., et al.: Generative AI for education (GAIED): advances, opportunities, and challenges. CoRR abs/2402.01580 (2024). https://doi.org/10.48550/ARXIV.2402.01580
5. Erola, J., Kilpi-Jakonen, E.: Family dynamics and child outcomes: an overview of research and open questions. Eur. J. Popul. **33**(2), 163–184 (2017)
6. Hanamanthrao, R., Thejaswini, S.: Real-time clickstream data analytics and visualization. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 2139–2144 (2017). https://doi.org/10.1109/RTEICT.2017.8256978
7. Käser, T., Schwartz, D.L.: Modeling and analyzing inquiry strategies in open-ended learning environments. Int. J. Artif. Intell. Educ. **30**(3), 504–535 (2020). https://doi.org/10.1007/S40593-020-00199-Y
8. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia Medica (Zagreb) **22**(3), 276–282 (2012). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/
9. Mirchandani, S., et al.: Large language models as general pattern machines. In: Conference on Robot Learning (2023). https://api.semanticscholar.org/CorpusID:259501163
10. Moore, E.B., Herzog, T.A., Perkins, K.K.: Interactive simulations as implicit support for guided-inquiry. Chem. Educ. Res. Pract. **14**(3), 257–268 (2013)
11. OpenAI: GPT-4o system card (2024). https://openai.com/index/gpt-4o-system-card/. Accessed 14 Nov 2024
12. Phung, T., et al.: Automating human tutor-style programming feedback: leveraging GPT-4 tutor model for hint generation and GPT-3.5 student model for hint validation. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, LAK 2024, Kyoto, Japan, March 18-22, 2024, pp. 12–23. ACM (2024). https://doi.org/10.1145/3636555.3636846,
13. Phung, T., et al.: Automating human tutor-style programming feedback: leveraging GPT-4 tutor model for hint generation and GPT-3.5 student model for hint validation. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, LAK 2024, Kyoto, Japan, March 18-22, 2024, pp. 12–23. ACM (2024). https://doi.org/10.1145/3636555.3636846
14. Radmehr, B., Singla, A., Käser, T.: PharmaSimText: a text-based educational playground filled with RL-LLM agents that work together even in disagreement. J. Educ. Data Min. **17**(1), 1–40 (2025). https://doi.org/10.5281/zenodo.14681290
15. Roll, I., Winne, P.H.: Understanding, evaluating, and supporting self-regulated learning using learning analytics. J. Learn. Analytics **2**(1), 7–12 (2015)
16. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A systematic survey of prompt engineering in large language models: Techniques and applications (2025). https://arxiv.org/abs/2402.07927
17. 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., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 165–179. Springer Nature Switzerland, Cham (2024)

18. Schulz, A., Voermanek, J.: How do help-seeking and help-abuse affect learning achievement in an interactive learning environment? Comput. Educ. Open **8**, 100247 (2025). https://doi.org/10.1016/j.caeo.2025.100247
19. Shved, E., Bumbacher, E., Mejia-Domenzain, P., Kapur, M., Käser, T.: Teaching and measuring multidimensional inquiry skills using interactive simulations. In: Artificial Intelligence in Education: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part I, pp. 482–496. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-64302-6_34
20. Swartout, W.R., Moore, J.D.: Explanation in second generation expert systems. In: David, J.M., Krivine, J.P., Simmons, R. (eds.) Second Generation Expert Systems, pp. 543–585. Springer, Berlin Heidelberg, Berlin, Heidelberg (1993)
21. Zhang, N., Biswas, G., Hutchins, N.: Measuring and analyzing students' strategic learning behaviors in open-ended learning environments. Int. J. Artif. Intell. Educ. **32**(4), 931–970 (2022). https://doi.org/10.1007/S40593-021-00275-X
22. Zimmerman, B.J.: Becoming a self-regulated learner: an overview. Theory Into Pract. **41**(2), 64–70 (2002). https://doi.org/10.1207/s15430421tip4102_2

Effects of Human and Automated Feedback on Pre-service Teachers' Nonverbal Teaching Skills During Microteaching

Su Mu[1], Xiaoran Chen[2], Xueni Liu[2], Yuyin Zhang[2], Yingbin Zhang[1(✉)], and Xiaoyong Hu[1,3(✉)]

[1] Institute of Artificial Intelligence in Education, South China Normal University, Guangzhou 510631, China
{musu,zyingbin,huxiaoy}@m.scnu.edu.cn
[2] School of Information Technology in Education, South China Normal University, Guangzhou 510631, China
{2023010079,liuxueni,2023020891}@m.scnu.edu.cn
[3] Center for Faculty Development, South China Normal University, Guangzhou 510631, China

Abstract. Teachers' nonverbal communication, such as facial expressions, posture, and paralanguage, significantly influences students' motivation and academic outcomes. These elements, collectively referred to as nonverbal teaching skills in this study, are essential for pre-service teachers to master. Although personalized feedback on such skills enhances skill development during microteaching, providing timely personalized feedback is labor-intensive for teacher educator. Artificial intelligence-based automated feedback offers a scalable solution, yet its effectiveness compared to traditional human feedback remains underexplored. This study addresses this gap through a randomized controlled experiment involving 67 pre-service teachers in Mainland China, assigned to three conditions: human feedback ($n = 20$), automated feedback ($n = 21$), and a hybrid approach combining both ($n = 26$). Five indicators of nonverbal teaching skills were analyzed across four teaching sessions. Key findings revealed that: (1) all groups exhibited an increased frequency of posture changes and spent more time facing forward across sessions; (2) the human feedback group experienced a reduction in facial expression variability, while the other groups showed no changes; and (3) no significant changes occurred in speech disfluency or positive facial expressions in any group. These results underscore artificial intelligence-based feedback as a viable complement to human feedback in teacher education.

Keywords: Nonverbal Teaching Skill · Automated Feedback · Microteaching · Teacher Education

1 Introduction

Teaching skills refer to the knowledge, decision-making, and actions that teachers employ to foster student learning effectively [1]. Researchers have proposed various taxonomies of teaching skills, ranging from very general (e.g., classroom management) to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 103–110, 2025.
https://doi.org/10.1007/978-3-031-99267-4_13

very specific skills (e.g., speech fluency) [2]. Different taxonomies are suited to different purposes. The current study focuses on teachers' proficiencies in utilizing paralanguage (e.g., volume and fluency), facial expression, and postures, which are referred to as nonverbal teaching skills. In Mainland China, the importance of these skills is emphasized by governmental professional competency standards for pre-service teachers [3]. Studies have found that these skills influence students' motivation and learning outcomes [4–6] and are as important as more complex skills, such as questioning, explanation, and demonstration [2].

Artificial intelligence (AI)-based automated feedback has shown promise in supporting the development of teaching skills [7, 8]. However, it remains unclear whether automated feedback improves pre-service teachers' nonverbal teaching skills. To address these gaps, this study conducted a randomized experiment to evaluate the impact of automated feedback provided by the Intelligent Training System for Basic Teaching Skills (ITSBTS), a tool for microteaching. We investigated two research questions (RQs):

- RQ1: Do pre-service teachers' nonverbal teaching skills change when receiving automated feedback during microteaching?
- RQ2: Do groups with different feedback sources (human feedback, automated feedback, and a combination of both feedback sources) vary in the changes?

2 Literature Review

2.1 Training Pre-service Teachers' Teaching Skills

Generally, there are two types of training activities for pre-service teachers' teaching skills: microteaching and teaching practicum [9]. Teaching practicum places pre-service teachers in real classroom settings, allowing them to apply teaching skills under the guidance of experienced educators. In contrast, microteaching refers to delivering brief lessons in an environment without students, or with a small group of peers [10] or simulated students in digital screens (simulation-based teaching) [9]. After the practice, pre-service teachers discuss their teaching behaviors with peers and instructors to identify areas for improvement. Different microteaching procedures exist, but researchers generally agree that feedback on microteaching sessions is critical for pre-service teachers' skill development [11, 12].

2.2 AI-Based Automated Feedback on Teaching Skills

Human-based feedback in microteaching is often limited by time and location, and may lack specificity [12, 13]. Moreover, in higher education, the large number of pre-service teachers makes it difficult for teacher educators to provide real-time and personalized feedback [14], thus decreasing the effectiveness of microteaching. Therefore, studies have explored AI-based automated feedback for teacher education and professional development [15]. For instance, [16] combined eye-tracing, video, audio, and accelerometer data to generate orchestration graphs that visualized teaching activities over time. [17] designed an AI-based system that fused RGB video and skeleton data to recognize teaching behavior and support teacher reflection.

AI-based automated feedback has been found to enhance teachers' skills in diagnostic reasoning about students' thinking [8, 13], questioning [7], and uptake of student contributions [18]. Nevertheless, automated feedback specifically designed for teaching skill development remains limited [19], and studies evaluating the impact of the automated feedback are fewer. Furthermore, the impact of automated feedback on pre-service teachers' nonverbal teaching skills is unclear.

3 The Intelligent Training System for Basic Teaching Skills

ITSBTS is a commercial tool developed by the educational technology company Zonekey[1]. Similar to traditional microteaching systems, ITSBTS uses microphones and cameras to collect audio and video data from microteaching sessions. Additionally, it employs several artificial intelligence toolboxes (e.g., OpenCV[2] and PaddleSpeech[3]) to analyze the audio and video data and extract 18 indicators of nonverbal teaching skills. These indicators encompass four dimensions of nonverbal teaching skills: paralanguage, facial expressions, posture, and blackboard writing (see Table 1 for example indicators), aligned with teachers' professional competency standards in Mainland China [3]. For example, ITSBTS uses OpenCV to capture teachers' body keypoints from video frames, which are then processed to classify body postures into predefined categories. The postures of all video frames in a teaching session are converted into indicators, such as posture change frequency and the tendency to face forward. ITSBTS visualizes the extracted indicators (see Fig. 1) and generates downloadable reports on teaching sessions, providing formative feedback on nonverbal teaching skills.



Fig. 1. A section of the visualization of skill indicators in ITSBTS (left panel) and a portion of the downloadable report (right panel).

[1] https://www.zonekey.com.cn/solutions/detail116.html.
[2] https://github.com/opencv/opencv?tab=readme-ov-file.
[3] https://github.com/PaddlePaddle/Paddle.

4 Methods

4.1 Participants

Participants were 78 pre-service teachers majoring in Educational Technology at a university in southern China. Of these, 47 were female and 31 were male. Sixty-three participants were freshmen, ten were sophomores, and five were juniors. None of the participants had received microteaching training prior to the experiment.

4.2 Experimental Design and Procedures

Recall that ITSBTS provided feedback on 18 indicators of nonverbal teaching skills. However, a pilot study, which involved 14 pre-service teachers (none of whom participated in the main experiment) and two partner experts in the field of microteaching, indicated a need for clear evaluation criteria to interpret the feedback (e.g., defining what constitutes good performance for each indicator). After discussions with them, we identified four indicators for which the criteria were relatively clear (the first four rows in Table 1), and they were selected for training in the experiment. Although it was difficult to decide the ideal range for posture change frequency, but the experts indicated that most pre-service teachers' body were relatively stiff during microteaching and had a low posture change frequency. Therefore, the experiment also included this indicator.

Table 1. The Five Indicators of Nonverbal Teaching Skills Selected in the Experiment.

Indicator	Description and evaluation criteria	Range
Disfluency	The average number of hesitations (e.g., "eh"), repetitions (e.g., "then… Then…"), and fillers (e.g., "well") per minute. Smaller values indicate better performance	≥0
Positive facial expressions	The proportion of happy facial expressions. Higher values indicate better performance	0% ~ 100%
Facial expression variety	The number of unique facial expressions in a teaching session. Five categories: happy, neutral, surprise, sad, and angry. Higher values indicate better performance	1 ~ 5
Tendency to face forward	The proportion of time where teachers are facing the position where students should be in. Higher values are better performance	0% ~ 100%
Posture change frequency	The average number of posture (including gestures) changes per minute. Ideal performance should be a medium frequency, neither too high nor too low	≥0

Before starting the microteaching training, all participants were asked to design a 10-min lesson (excluding time for interactions and activities). The domain was chosen by

the participants, but the target audience had to be K-12 students. Four graduate research assistants, enrolled in a master's program in teacher education with teaching experience, provided suggestions on the initial lesson designs, which participants then revised. Participants then presented the revised lessons four times in a microteaching classroom equipped with ITSBTS, once per week. Immediately after each presentation (teaching session), participants received feedback on the five indicators in Table 1 from different sources, based on their randomly assigned group membership: for the *human feedback (HF) group*, graduate research assistants provided feedback based on their observations of participants' teaching sessions; the *automated feedback (AF) group* received feedback from the ITSBTS reports; for the *human + automated feedback (HAF) group*, graduate research assistants provided feedback based on both their observations and the ITSBTS reports. Participants were asked to reflect on their teaching sessions based on the feedback received and attempt to make improvements in the next session.

4.3 Analytic Approaches

We excluded data from 11 participants who presented only one teaching session due to personal reasons. The final group sizes were 20 for the HF group, 21 for the AF group, and 26 for the HAF group. The Shapiro-Wilk tests indicated that the five indicators did not follow a normal distribution in most cases ($p < 0.05$). Therefore, we applied a two-way, rank-based repeated ANOVA using the R package nparLD [20] to evaluate the impact of feedback sources on participants' nonverbal teaching skills. Time (1st to 4th teaching sessions) was the within-subject factor, and group membership (HF, AF, and HAF) were the between-subject factor. The repeated ANOVA produced a statistic that could be approximated by an F-distribution with infinite degrees of freedom in the denominator [21].

5 Results

Figure 2 displays the raw values of the five indicators of nonverbal teaching skills. The non-parametric repeated ANOVA indicated that for disfluency and positive facial expressions, neither the main effects of time nor the interactions between time and group were statistically significant ($p \geq 0.17$ for all effects). The results suggest that no changes occurred in disfluency and positive facial expressions across all groups over time. For expression variety, the main effect of time was not significant ($F_{2.92,\infty} = 1.45, p = 0.23$, partial $\eta^2 = 0.04$), but the interaction between time and group was significant ($F_{5.35,\infty} = 3.02, p = 0.01$, partial $\eta^2 = 0.14$), suggesting that the groups differed in how the number of unique facial expressions changed over sessions. Simple effect analyses (with Holm–Bonferroni correction) revealed that the HF group's expression variety changed over time ($p = .004$), while the other groups did not ($p \geq 0.41$). According to Fig. 2, the HF group's expression variety decreased from Time 3 to Time 4.

For posture change frequency, the main effect of time was significant ($F_{2.47,\infty} = 5.45$, $p < 0.01$, partial $\eta^2 = 0.09$), but the interaction between time and group was not significant ($F_{4.50,\infty} = 1.03$, $p = 0.39$, partial $\eta^2 = 0.08$). Participants' posture change frequencies increased over time (see Fig. 2), and groups had no difference in

the frequency increases. For the tendency to face forward, the main effect of time was significant ($F_{2.53,\ \infty} = 7.77, p < 0.01$, partial $\eta^2 = 0.16$), but the interaction between time and group was not significant ($F_{4.64,\ \infty} = 0.88, p = 0.49$, partial $\eta^2 = 0.05$). Participants' tendency to face forward increased over time (see Fig. 2), with no difference between groups.

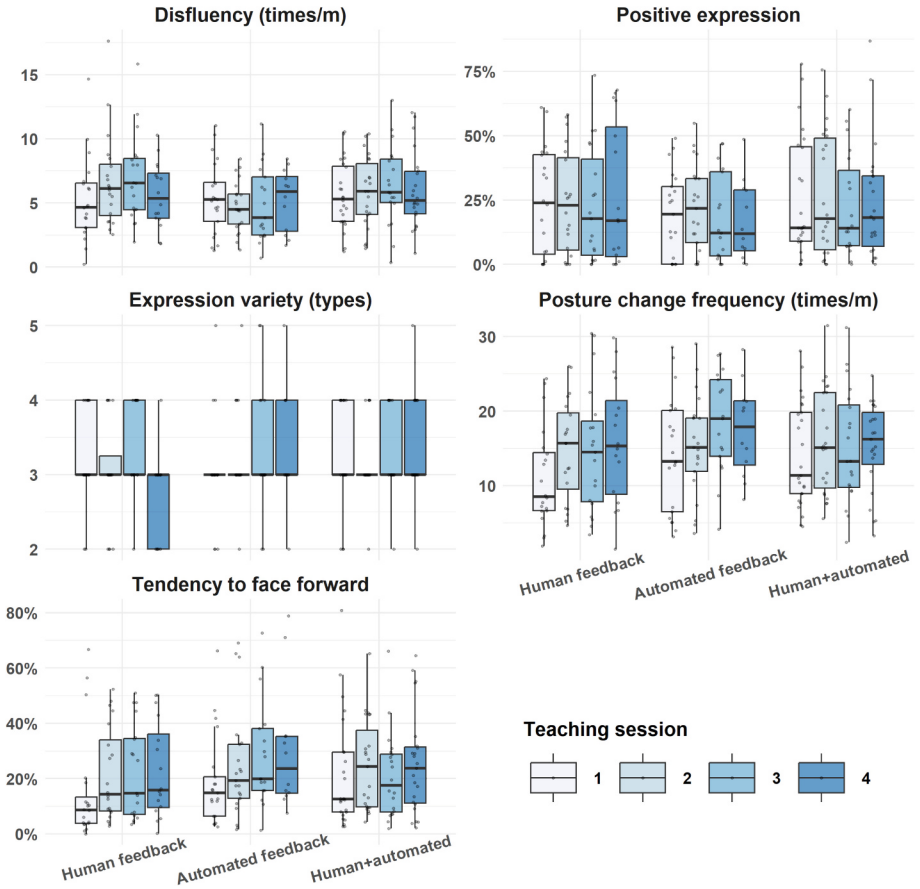


Fig. 2. The five indicators of nonverbal teaching skills in each group at each teaching session.

6 Conclusions, Limitations, and Future Directions

For RQ1, the results indicated that, same as those with human feedback, the AF group's posture change frequency and tendency to face forward increased after four teaching sessions, while the other indicators did not change. For RQ2, the HF group showed a decrease in facial expression variety, while the other groups did not. Overall, the results suggest that automated feedback may improve certain indicators of pre-service teachers'

nonverbal teaching skills, with effects comparable to human feedback. Combining both feedbacks may not provide any additional benefit over automated feedback alone.

This study has several limitations. It only evaluated five indicators of nonverbal teaching skills due to the lack of clear evaluation criteria for the other indicators. Future research should develop such criteria and evaluate the impact of automated feedback on additional indicators. The ITSBTS report included indicators beyond the five selected for the current study. Students might focus on indicators they felt performed poorly, potentially neglecting the five selected indicators for improvement. All participants were majoring in Educational Technology. Their AI knowledge and attitude toward automated feedback, which may influence the effect of automated feedback [22], could differ from that of pre-service teachers in other majors, limiting the generalizability of the results. Additionally, most participants were freshmen with limited teaching experience, which may constrain the generalizability of the findings to individuals with more experience, as experience is likely an important factor influencing performance.

Acknowledgments. This study was funded by National Natural Science Foundation of China (No.62407014).

References

1. Kyriacou, C.: Essential Teaching Skills. 5th ed. Oxford University Press-Children (2018)
2. Zhang, J., Pan, Q., Zhang, D., Meng, B., Hwang, G.: Effects of virtual reality based microteaching training on pre-service teachers' teaching skills from a multi-dimensional perspective. J. Educ. Comput. Res. **62**(3), 875–903 (2024)
3. Ministry of Education, the People's Republic of China: Notice of the General Office of the Ministry of Education on issuing five documents including the "Professional competency standards for secondary education teacher trainees (Trial). Ministry of Education Government Portal (2021). http://www.moe.gov.cn/srcsite/A10/s6991/202104/t20210412_525943.html
4. Traulsen, S.J., Zander, L.: Straighten your back, open your arms! Effects of instructor's body postures in educational videos on students' interest and motivation. Learn. Instr. **93**, 101959 (2024)
5. Wang, Y., Liu, Q., Chen, W., Wang, Q., Stein, D.: Effects of instructor's facial expressions on students' learning with video lectures. Brit. J. Educ. Technol. **50**(3), 1381–1395 (2019)
6. Paulmann, S., Weinstein, N.: Teachers' motivational prosody: a pre-registered experimental test of children's reactions to tone of voice used by teachers. Brit. J. Educ. Psychol. **93**(2), 437–452 (2023)
7. Demszky, D., Liu, J., Hill, H.C., Sanghi, S., Chung, A.: Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: opportunities for further enhancement. Comput. Educ. **227**, 105183 (2025)
8. Bauer, E., et al.: AI -based adaptive feedback in simulations for teacher education: an experimental replication in the field. J. Comput. Assist. Lear. **41**(1) (2025)
9. Ledger, S., Fischetti, J.: Micro-teaching 2.0: technology as the classroom. Australas. J. Educ. Tec. **36**(1), 37–54 (2019)
10. Allen, D.W., Cooper, J.M., Poliakoff, L.: Microteaching. US Department of Health, Education, and Welfare (1972)
11. Wang, M., Long, T., Li, N., Shi, Y., Chen, Z.: The impact of different types of feedback on pre-service teachers' microteaching practice and perceptions. Educ. Inf. Technol. (2024)

12. Prilop, C.N., Weber, K.E., Kleinknecht, M.: The role of expert feedback in the development of pre-service teachers' professional vision of classroom management in an online blended learning environment. Teach. Teach. Educ. **99**, 103276 (2021)
13. Sailer, M., et al.: Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. Learn. Instr. **83**, 101620 (2023)
14. Powell, C.G., Bodur, Y.: Teachers' perceptions of an online professional development experience: implications for a design and implementation framework. Teach. Teach. Educ. **77**, 19–30 (2019)
15. Celik, I., Dindar, M., Muukkonen, H., Järvelä, S.: The promises and challenges of artificial intelligence for teachers: a systematic review of research. TechTrends **66**(4), 616–630 (2022)
16. Prieto, L.P., Sharma, K., Kidzinski, A., Rodríguez-Triana, M.J., Dillenbourg, P.: Multimodal teaching analytics: automated extraction of orchestration graphs from wearable sensor data. J. Comput. Assist. Lear. **34**(2), 193–203 (2018)
17. Luo, H., et al.: The recognition of teacher behavior based on multimodal information fusion. Math. Probl. Eng. **2020**(2020), 1–8 (2020)
18. Demszky, D., Liu, J., Hill, H.C., Jurafsky, D., Piech, C.: Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. Educ. Eval. Policy An. **46**(3), 483–505 (2023)
19. Copur-Gencturk, Y., Li, J., Cohen, A.S., Orrill, C.H.: The impact of an interactive, personalized computer-based teacher professional development program on student performance: a randomized controlled trial. Comput. Educ. **210**, 104963 (2024)
20. Noguchi, K., Gel, Y.R., Brunner, E., Konietschke, F.: NparLD: an r software package for the nonparametric analysis of longitudinal data in factorial experiments. J. Stat. Softw. **50**(12) (2012)
21. Brunner, E., Puri, M.L.: Nonparametric methods in factorial designs. Stat. Pap. **42**(1), 1–52 (2001)
22. Scherer, R., Tondeur, J., Siddiq, F., Baran, E.: The importance of attitudes toward technology for pre-service teachers' technological, pedagogical, and content knowledge: comparing structural equation modeling approaches. Comput. Hum. Behav. **80**, 67–80 (2018)



Avalon: A Human-in-the-Loop LLM Grading System with Instructor Calibration and Student Self-assessment

Derek Armfield[1], Eason Chen[1(✉)], Asilbek Omonkulov[1], Xinyi Tang[1], Jionghao Lin[2], Erik Thiessen[1], and Kenneth Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
eason.tw.chen@gmail.com
[2] The University of Hong Kong, Hong Kong, China

Abstract. Open-ended assessments promote deeper learning but pose significant challenges for timely, high-quality feedback–often burdening instructors with lengthy grading processes and resulting in underutilized student feedback. We introduce Avalon, a human-in-the-loop AI grading system that integrates (1) an iterative calibration phase to align AI graders with instructor expectations and (2) a novel student self-grading and discrepancy reporting mechanism. Through rubric calibration, instructors provide corrective feedback on AI-graded samples, ensuring consistent application of grading criteria. After the AI grades all submissions, students assess their own work using the same rubric, then compare their scores with the AI's. They submit short "discrepancy reports" for any mismatches, distinguishing between accepted differences and genuine disputes. In a pilot with 102 undergraduates, Avalon reduced instructor grading time by focusing manual review on a small subset - fewer than 16% of submissions - that were disputed by students. Moreover, students show high engagement with the feedback process, as self-grading compelled them to revisit rubric criteria and reflect on their submissions more deeply. Furthermore, the system uncovered misconceptions that might otherwise have gone undetected–prompting targeted instructor intervention. Although additional validation and larger-scale studies are needed, current preliminary findings suggest Avalon's hybrid approach can reduce grading workloads, improve feedback effectiveness, and enhance student engagement and learning outcomes. The Avalon platform can be accessed at https://avalonlearn.com

Keywords: Large Language Models · Grading System · Feedback

1 Introduction and Related Works

Large enrollment courses and online learning platforms (e.g., MOOCs) [1] offer unprecedented access to education, but they also place immense grading burdens on instructors [6]. Open-ended assignments–ranging from short answers to lengthy essays or proofs–are particularly time-consuming to grade [11,13]. Ironically, despite the rich feedback instructors give, many students do not fully read

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 111–118, 2025.
https://doi.org/10.1007/978-3-031-99267-4_14

or act on these comments [5,16,18,23]. Prior research shows that only 58% of feedback gets accessed [16], implying lost opportunities for deeper learning.

Automated grading systems promise to reduce grading overhead [17] and deliver timely and personalized feedback [3,14], especially with recent advances in large language models (LLMs) [2,4,7,8]. However, completely delegating grading to AI raises concerns over accuracy, transparency, and accountability [10,22]. Even though researchers report that AI-generated grades can align human scores moderately well, doubts persist regarding whether AI can consistently evaluate nuanced student thinking [8]. These concerns highlighted the need for a *hybrid* approach that combines LLM capabilities with instructor oversight and more engaging mechanisms for students to receive feedback and raise disputes.

Recent research has shown that LLMs can achieve high agreement with human instructors on short-answer and essay grading [8,9], yet fully automated systems often lack transparency, ignore student agency, and cannot detect shared misunderstandings between AI and student. Recent hybrid approaches involving instructorâĂŞAI collaboration (e.g., rubric refinement [19]) and AI-assisted code grading with TA verification [20] show promise, but typically omit student involvement. Similarly, self-assessment tools have been shown to improve metacognition [12,21], but are rarely integrated with AI grading.

To address these gaps, we propose Avalon, a human-in-the-loop AI grading system that combines instructor calibration, large language model scoring, and structured student self-assessment. Rather than treating evaluation as something that happens to the student and their work, Avalon reimagines grading as a collaborative process with the student–transforming them from a passive consumer of feedback and evaluation into a key, active participant who both deepens their engagement and strengthens the accuracy of assessment.

We piloted Avalon in an undergraduate course with 102 students and 507 submissions. *Preliminary findings in the pilot study* indicate that

(a) **78% of students engage in receiving feedback**, as Avalon encourages them to revisit rubric criteria for self-grading, then resolve misalignment with the AI grader–whether by disputing or reflecting on their grading results.
(b) An estimated **reduction in grading time of over 85%** since the instructor only needs to review the fewer than 16% of submissions which were disputed.
(c) **Only 9 out of 1729 of the AI's judgements were overturned** by the instructor due to AI misinterpretation, suggesting strong alignment with the instructor's criteria.

These promising outcomes highlight Avalon's potential to ease instructor workloads while fostering more meaningful student feedback utilization.

2 Avalon System Implementation

Figure 1 illustrates how instructors and students engage with the grading agent and self-grading system. Specifically, they will follow the following steps:

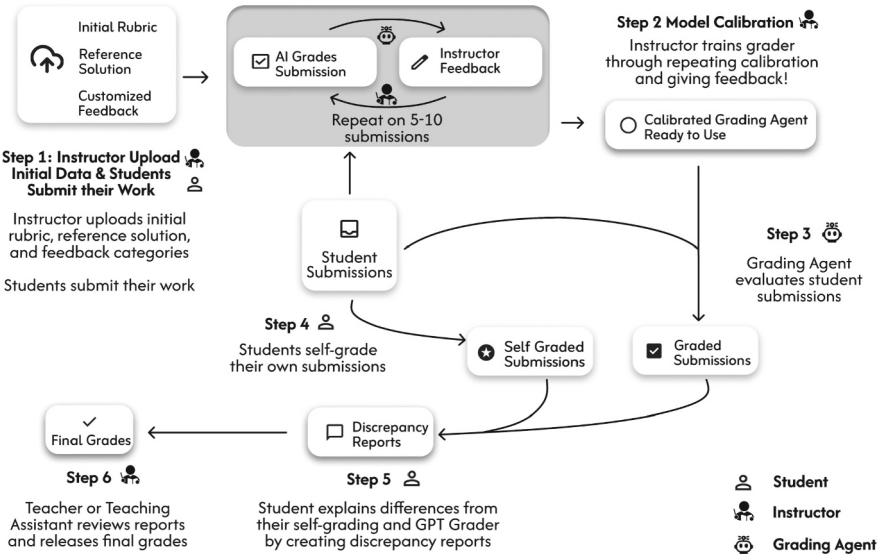


Fig. 1. User flow for Avalon. Please refer Sect. 2 for explanation of each step.

Step 1: Instructor Uploads Material and Students Submit Their Work. In order to initiate the grading process, several components must be in place. First, the instructor creates a preliminary rubric consisting of concise item descriptions paired with associated point values. This rubric serves as a tentative hypothesis for how the assignment should be graded and will be refined throughout the calibration procedure. Additionally, the instructor provides a reference solution, which offers the Grading Agent context regarding the expected solution. The instructor may also define custom feedback categories—for example, *strengths*, *weaknesses*, and *improvements*—or adapt these categories to address specific learning objectives. Crucially, students must submit their work so the instructor can begin the calibration process; however, it is possible to do the calibration process using only a subset of the total submissions.

Step 2: Rubric Calibration. Rubric calibration ensures that the Grading Agent, powered by GPT-4o, interprets the rubric in accordance with the instructor's intended grading criteria via iterative prompt-engineering. This procedure, inspired by human grading practices often referred to as *normalization*, proceeds one submission at a time. Both the Grading Agent and the instructor independently evaluate the same submission. Whenever a discrepancy arises, the instructor provides corrective feedback. The Grading Agent then proposes adjustments to the rubric to prevent future misinterpretations. The instructor may accept this suggestion or modify the rubric further to align with their vision for how the assignment should be graded. This iterative process continues until

the instructor is satisfied that the Grading Agent is consistently applying the rubric in agreement with the instructor.

Step 3: Grading Agent Evaluates All Submissions. Once calibration is complete, the instructor has gained sufficient confidence that the Grading Agent can reliably apply the rubric. At this stage, the Grading Agent proceeds to evaluate all remaining student submissions. Points are awarded, and feedback is generated based on the finalized rubric.

Step 4: Students Submit Their Self-Grading. After the Grading Agent has completed its evaluations, students are given the opportunity to grade their own work using the same rubric. Notably, they do not see the Grading Agent's assigned scores or feedback beforehand. This self-assessment requires students to interpret and apply the rubric to their own submissions, encouraging deeper engagement with the criteria used to evaluate their work.

Step 5: Student View Results and Create Discrepancy Report. Following the self-grading phase, the Grading Agent's evaluation is revealed. The system compares the student's self-assessment with the agent's grading to identify any discrepancies for each rubric item. The student is encouraged to submit a *discrepancy report* for each identified mismatch. Two primary types of discrepancy reports occur:

1. **Self-Acknowledged**: The student concedes that their initial self-assessment was incorrect. In such cases, the student provides a brief explanation detailing why their original understanding differed from the Grading Agent's evaluation and clarifies what led to their revised perspective.
2. **Disputed**: The student maintains that their original assessment was correct, thus contesting the Grading Agent's evaluation. Here, the student argues why they believe the rubric item should (or should not) have been applied as initially self-assessed. Although rare, it is possible for a student to argue against awarding themselves positive points if they believe the Grading Agent misapplied the rubric.

Step 6: Instructor Review Reports and Release Final Grades. Finally, the instructor and teaching assistants review the submitted discrepancy reports. While it may be sufficient to focus on the *disputed* reports, instructors can also examine *self-acknowledged* reports to gain insight into common student misunderstandings and adapt their teaching strategy in the future, such as emphasizing the misunderstanding part. When necessary, the instructor can adjust scores or offer clarifications. After any revisions are finalized, the instructor releases the official grades to the students.

3 Discussion with Preliminary System Evaluation Results

Preliminary system evaluation with 102 students from an undergraduate course at a university suggests several promising outcomes. Our study was approved by the IRB. This section highlights three key observations: (1) reductions in grading time, (2) improved delivery for feedback (3) high agreement for the AI grading result, and (4) increased student engagement facilitated by self-grading and self-reflection on potential misconceptions.

(1) Grading Time Reduction. In total, 77.12% of all submissions received were self-graded by students. Among these, 54.48% (equivalent to 42.01% of all submissions) exhibited differences in at least one rubric item compared to the AI grader (note that each submission may have more than one rubric item with discrepancies). Among these differences, 66.20% (27.81% overall) led to the submission of a discrepancy report. Within these reports, 56.74% contained at least one disputed rubric item, implying that only 15.78% of all submissions ultimately required direct instructor intervention.

Because instructors only have to review submissions with a disputed rubric item (approximately 15%), Avalon can potentially reduce instructor grading time by around 85%. In other words, an assignment that would normally take 20 h of human grading time could now be finished in less than 3 h. This estimated reduction may even be conservative since instructors typically only have to resolve the disputed rubric items – often just one out of four rubric items – and resolving these reports was reportedly 2–3x faster (from instructor interview) than grading from scratch. That same 20 h of human grading time could be reduced to just 1 h. Since calibration's impact on overall grading time is minimal as it does not scale with class size, is already relatively quick (under one hour), and can be expected to become even faster as instructors gain familiarity with the system, it has been omitted from these estimates.

(2) Improved Deliverability of Feedback. Previous studies have shown that up to 42% of feedback goes unaccessed when grades can be viewed elsewhere [16]. In contrast, our system requires students to complete a self-grading exercise before viewing any instructor or AI-generated comments. This additional step increases feedback accessed: only 22.28% of the feedback goes unaccessed despite the extra effort required from students.

(3) High Agreement, and Misconception Detection. We analyzed all submissions that were graded both by the Grading Agent and by the students themselves. These comprised five questions, a total of 391 submissions, which include 1332 rubric items assessed by both the AI system and the students. We observed an overall alignment of 78% between the student-self-grading and AI Agent's scores on rubrics. Then, when students compared their self-assessments with the AI's evaluations, we observed that 93% of student evaluations agreed

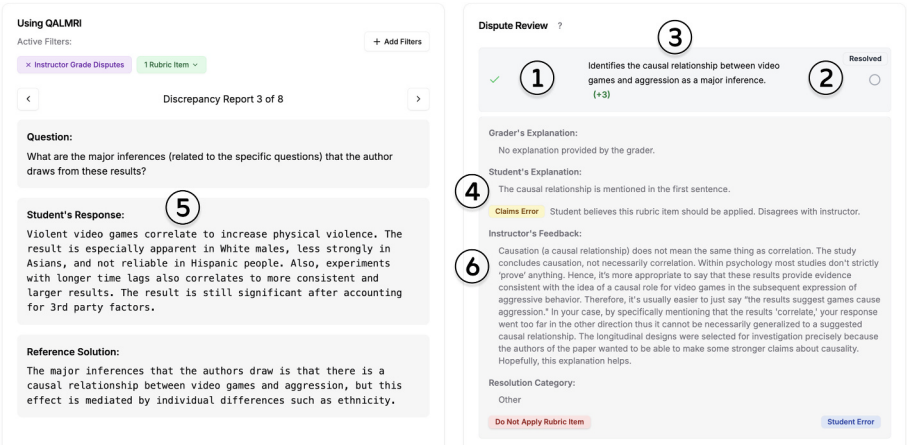


Fig. 2. A screenshot from Avalon's interface illustrates a student's misunderstanding. The green check (1) shows the student believed the rubric item should be applied, whereas the grey circle (2) indicates the Grading Agent did not apply it. Despite this, the student argued their response met the rubric item's criteria (3) of demonstrating a causal relationship as evidenced by their first sentence (4), although it actually only mentioned a correlation (5). The instructor resolved the dispute by not applying the rubric item and providing targeted feedback (6), highlighting how such misunderstandings can be identified and clarified.

with the AI's judgment, resulting in a Kappa of 0.84, indicating strong agreement [15]. That being said, it is possible that both the student and the AI are wrong while in agreement. Future work would then undergo a human review for all submissions to mitigate this issue.

Although some students still disagreed with AI's grading and submitted discrepancy reports to dispute it, most of the discrepancy reports were ultimately attributed to student misunderstandings of the rubric by the instructor, like the example illustrated in Fig. 2. Out of 1729 AI grading judgments, only 9 were overturned due to AI misinterpretation. This result underscores a high level of consistency between the AI grader and the instructor's assessments.

Furthermore, the self-grading component appeared pivotal in uncovering misconceptions and promoting engagement. By requiring students to explicitly assess their own work using the same rubric, this system prompted students to reflect on their reasoning processes and confront any incongruities with the AI's evaluations. Instructors could then focus their efforts on resolving the most substantive disputes, where students genuinely misunderstood the material or rubric criteria, thus enabling more targeted and impactful teaching and feedback.

4 Conclusion

In this paper, we introduce *Avalon* – a platform that seeks to rethink the student's role in feedback and evaluation, reduce time spent on grading open-ended

assignments, and provide consistency, transparency, and interpretability to AI grading systems. By having students independently self-grade their own submissions, comparing this self-grade to the AI grading, submitting short explanations on why any item may differ, and then having instructors respond to such reports with tailored feedback, *Avalon* manages to drastically increase student engagement, reflection, and interaction with feedback. The iterative calibration process, having the agent grade and the instructor providing corrective feedback, helps align the AI Grading Agent to human grading standards.

References

1. Baturay, M.H.: An overview of the world of MOOCs. Procedia. Soc. Behav. Sci. **174**, 427–433 (2015)
2. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
3. Chen, E., Lee, J.E., Lin, J., Koedinger, K.: GPTutor: great personalized tutor with large language models for personalized learning content generation. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale, pp. 539–541 (2024)
4. Chen, E., Wang, D., Xu, L., Cao, C., Fang, X., Lin, J.: A systematic review on prompt engineering in large language models for k-12 stem education. arXiv preprint arXiv:2410.11123 (2024)
5. Crisp, B.R.: Is it worth the effort? How feedback influences students' subsequent submission of assessable work. Assess. Eval. High. Educ. **32**(5), 571–581 (2007)
6. Dunlap, J.C.: Workload reduction in online courses: getting some shuteye. Perform. Improv. **44**(5), 18–25 (2005)
7. Figueras, C., Farazouli, A., Cerratto Pargman, T., McGrath, C., Rossitto, C.: Promises and breakages of automated grading systems: a qualitative study in computer science education. Educ. Inq. 1–22 (2025)
8. Flodén, J.: Grading exams using large language models: a comparison between human and AI grading of exams in higher education using ChatGPT. Br. Educ. Res. J. (2024)
9. Gobrecht, A., et al.: Beyond human subjectivity and error: a novel AI grading system. arXiv preprint arXiv:2405.04323 (2024)
10. Idris, M.D., Feng, X., Dyo, V.: Revolutionising higher education: Unleashing the potential of large language models for strategic transformation. IEEE Access (2024)
11. Jerrim, J., Sims, S.: Teacher workload and well-being. new international evidence from the OECD TALIS study. Teaching and Teacher Education (2020)
12. Khojasteh, L., Kafipour, R., Pakdel, F., Mukundan, J.: Empowering medical students with AI writing co-pilots: design and validation of AI self-assessment toolkit. BMC Med. Educ. **25**, 159 (2025). https://doi.org/10.1186/s12909-025-06753-3
13. Kreuzfeld, S., Felsing, C., Seibt, R.: Teachers' working time as a risk factor for their mental health-findings from a cross-sectional study at German upper-level secondary schools. BMC Public Health **22**(1), 307 (2022)
14. Lin, J., et al.: How can i improve? Using GPT to highlight the desired and undesired parts of open-ended responses. arXiv preprint arXiv:2405.00291 (2024)
15. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)

16. Mensink, P.J., King, K.: Student access of online feedback is modified by the availability of assessment marks, gender and academic performance. Br. J. Edu. Technol. **51**(1), 10–22 (2020)
17. 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)
18. Murtagh, L., Baker, N.: Feedback to feed forward: student response to tutors' written comments on assignments. Pract. Res. High. Educ. **3**(1), 20–28 (2009)
19. Myint, P., Lo, S.L., Zhang, Y.: Harnessing the power of AI-instructor collaborative grading approach: topic-based effective grading for semi open-ended multipart questions. Comput. Educ. Artif. Intell. **7**, 100339 (2024). https://doi.org/10.1016/j.caeai.2024.100339
20. Nagakalyani, G., Chaudhary, S., Apte, V., Ramakrishnan, G., Tamilselvam, S.: Design and evaluation of an AI-assisted grading tool for introductory programming assignments: an experience report. In: Proceedings of the 56th ACM Technical Symposium on Computer Science Education (SIGCSE) (2025)
21. Nieminen, J.H., et al.: Self-assessment design in a digital world: centring student agency. Assess. Eval. High. Educ. (2025). https://doi.org/10.1080/02602938.2025.2467647
22. Quttainah, M., Mishra, V., Madakam, S., Lurie, Y., Mark, S., et al.: Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education: narrative review and qualitative study. JMIR AI **3**(1), e51834 (2024)
23. Woods, D.: Students viewing of feedback: an exploration of technology-mediated learning. J. Educ. Technol. Syst. **51**(1), 46–62 (2022)

Empowering Teachers with Technology: A National Study on a Formative Assessment Platform

Mingyu Feng(✉), Linlin Li, Chunwei Huang, Natalie Brezack, and Kim Luttgen

WestEd, San Francisco, CA 94107, USA
{mfeng,lli,chuang,nbrezac,kluttge}@wested.org

Abstract. Boosting math instruction and learning in the U.S. is a broadly recognized goal for math education, and the need is particularly strong for low-performing students and those from rural areas. A proven technology-based solution was recently augmented with a virtual professional learning community to overcome barriers for broad adoption and empower teachers. The platform, called "ASSISTments," provides immediate feedback to students as they solve math problems and automatically prepares reports for teachers about student performance on daily assignments. A nationwide experimental study involving 39 schools from 24 states across the U.S. examined the efficacy of the educational technology innovation targeting middle school students. The paper describes the sample, the study design, and the program's implementation in schools. Findings suggest that the innovation significantly improved low-performing students' math learning outcomes and can be widely implemented with adequate fidelity.

Keywords: Math learning · Formative Assessment · Teacher Professional Learning · AI-based Program · School Implementation · Learning Impact

1 Introduction

Mathematics education remains a critical focus for national education improvement in the U.S., particularly following learning loss due to the COVID-19 pandemic. In 2024, the NAEP math score for eighth graders dropped by 8 points compared to 2019 [1]. Independent practice and teachers' use of resulting data to inform instruction are crucial for student learning [2]. However, paper-based independent practice has limitations: when students complete their work, they may practice incorrectly, and teachers have limited time to analyze students' progress and adjust instruction [3].

Additionally, there is an acute need to target interventions to teachers and students in rural communities who face multifaceted challenges to accessing high quality math education [4]. Approximately one third of U.S. students attend rural schools, which comprise 53% of districts [5]. Teachers in rural schools may be the only teacher for miles who teach a subject area or grade level, leading to professional isolation. In-person professional learning options are limited by practical considerations; schools

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 119–126, 2025.
https://doi.org/10.1007/978-3-031-99267-4_15

are far apart, teachers face logistics challenges (e.g., finding substitutes), and there are limited available programs.

Educational technologies like ASSISTments provide opportunities to address both challenges. ASSISTments is a free, curriculum-based platform that provides independent practice for students and formative assessment analytics for teachers. The platform consists of an extensive content library of high-quality Open Educational Resource curricula [6] and standard-aligned problem sets of over 200,000 problems to support student learning and assess their understanding. As students work on problems, the system provides immediate feedback and offers hints or scaffolds. ASSISTments provides teachers with real-time reports that summarize student work; these data serve as evidence of proficiency to inform teachers to adjust instruction. The program was identified as effective in improving 7th grade student's learning at the end of the school year ([7][8]; $g = 0.22$, $p < 0.05$) and over the long term ([9][10]; $g = 0.10$, $p < 0.01$) and in changing teacher's review practices [7]. It is especially beneficial for students who started with lower prior knowledge ([7]; $g = 0.29$).

During prior state-wide studies, teachers travelled to a central location to in-person, intensive summer professional learning workshops to learn to use ASSISTments. Coaches conducted regular classroom visits to provide support up to three times per year. This model, while effective, became a barrier for scaling up the intervention nationwide, especially in rural districts where schools are far apart, have low enrollment levels, and may only have one middle school math teacher. The overhead of coaches traveling to each school becomes unaffordable and there is often no accessible central location for workshops. In a recent project, a virtual professional learning community (vPLC) was created to augment the ASSISTments platform and remove implementation barriers, allowing rural teachers to share best practices for data-driven math instruction across the nation in an online community via facilitated discussions. vPLCs channel the benefits of in-person sessions while making professional learning accessible [11][12]. In rural settings, vPLCs are effective when they are flexible, interactive, relevant, and provide ongoing support [13].

Supported by the U.S. Department of Education, we conducted a nationwide, large-scale experimental study to provide insight about the efficacy of ASSISTments with vPLC. This late-breaking results paper reports on the findings from the recently completed analysis of the study and address the following research questions:

- What is the effect of the vPLC-augmented ASSISTments platform on 6th - 8th grade students' math achievement, particularly for those in rural schools?
- Do the effects vary for students with different prior achievement?
- Do participating teachers implement the program as intended by the developer?

2 ASSISTments, the Formative Assessment Platform that Empowers Teachers

ASSISTments uses technology to give teachers new capabilities for assigning and reviewing homework and classwork and gives students support for learning. ASSISTments content consists of math problems with answers and hint messages bundled in problem sets which teachers can assign. Students do the problems on paper and enter

their answers into ASSISTments to receive immediate feedback about correctness, hints, or help separating multi-step problems into parts (Fig. 1). One type of problem set is mastery-oriented "Skill Builders," which provide practice with solving problems focusing on a targeted skill, until students reach a teacher-defined "mastery" threshold (e.g., three correct answers). ASSISTments also automatically re-assesses students on skills and concepts that were "mastered" at regular intervals, providing further opportunities for students to hone skills. ASSISTments provides teachers with real-time, easily accessible reports that summarize student work for an assignment in a grid format, which inform teachers about the average percent correct per question, skills covered, common wrong answers, and each student's answer to every question (Fig. 1). Teachers use data to form instructional groups and address common difficulties.

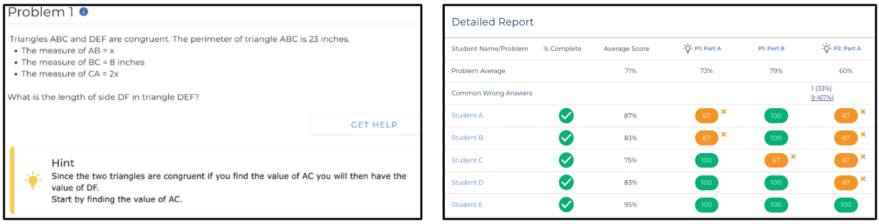


Fig. 1. Left: ASSISTments Student Interface; right: ASSISTments Teacher Detailed Report.

The vPLC addressed analysis-of-practice, an evidence-based method for improving learning [14]. It consisted of eight facilitated, synchronous sessions over the school year, approximately once per month. The sessions, orchestrated by a facilitator, gradually moved from training to support following a facilitation guide.[1] The topics included how to a) set high expectations for students, b) find appropriate content to assign and review reports to differentiate instruction, c) construct effective feedback and orchestrate open classroom discussion, and 4) create a supportive learning environment and motivate students with routines and reports. The sessions were held online over videoconference where screen-sharing and breakout rooms allowed for rich interactions. Over the school year, teachers followed four steps of formative assessment for implementing ASSISTments: 1) creating math assignments aligned with grade-level standards, 2) encouraging students to complete assignments and use immediate feedback and hints to support problem-solving, 3) obtaining reports that summarize individual and class performance and common wrong answers, and 4) analyzing the anonymous data with the class, discussing common errors, and customizing assignments.

ASSISTments was designed based upon theoretical foundations and research. *Feedback* has been identified as a powerful way to increase student learning [15]. An extensive set of studies (e.g., [3, 16, 17] has found learning gains in response to feedback within computer-based instruction. *Formative assessment* [18, 19] informs teachers of student learning and progress and informs students of their performance in relation to learning goals. Several decades of research have shown formative assessment to be an effective way of improving teaching and learning [20, 21]. *Mastery learning* has been

[1] The facilitation guide has been released to the public at https://info.assistments.org/plc-guide.

shown to lead to higher achievement than traditional forms of teaching [22, 23]. The design of vPLC also followed evidence and theory. Virtual PLCs promote greater self-reflection than in-person PLCs [12]. Greater teacher PLC engagement has been related to changes in instructional practices (differentiated instruction; [24]), and even teachers who participate with lower levels of engagement can still see benefits [25].

3 Research Design

This study[2] used a matched quasi-experimental design: treatment students using ASSISTments were compared with a virtual comparison group (VCG[3]) drawn from a national testing database, through Northwest Evaluation Association (NWEA), a K-12 assessment and research company and the provider of MAP Growth assessment to millions of students in the U.S. each year. In the fall, treatment students took the NWEA MAP Growth Math Assessment [27] as a baseline measure of math achievement and then took the same test in the spring as the posttest. For each student in the treatment group, NWEA identified matched non-treatment group students in their national database based on school locale, school-level percentage of students eligible for free- or reduced-priced lunch, test-taking dates, grade level, and fall baseline test scores. The mean of the posttest scores of up to 51[4] random samples of the matched students was considered a posttest score for the "virtual comparison" student that was paired with the treatment student and was used in estimating impact. This process occurred for every treatment student to form the comparison group.



Fig. 2. Geographic distribution of participating schools. Green indicates rural schools.

The study took place in two cohorts (2022–23 and 2023–24) involving a diverse sample of 36 schools across 24 U.S. states with 59 middle school teachers and their

[2] The study was approved by the Institutional Review Board at Worcester Polytechnic Institute. Participating teachers signed consent forms. Parents received consent forms to give permission for their children to participate. The study was pre-registered at the Registry of Efficacy and Effectiveness Studies, #8560.1v2.

[3] The VCG was developed by NWEA as an alternative to conventional controlled experiments. The procedures for randomly selecting cases for membership in a VCG create samples that do not differ significantly from the population group from which they are drawn [26].

[4] The technical details to determine the number of samples to aggregate was discussed in [26].

math classrooms. Figure 2 shows the wide geographically distribution of the schools. Eighteen schools (with 28 teachers) from 12 states were rural according to U.S. NCES locale categorization. More than half of the students (52%) were from low socioeconomic status backgrounds. Teachers tended to be from schools that had experienced a drop in math performance and were interested in boosting math learning using technology.

4 Data, Analysis and Findings

4.1 Data Collection

Outcome Measure. The study measured student math achievement with the online MAP Growth Math Assessment, which have demonstrated reliability and validity [28]. The test was administered in the fall as a baseline and in the spring as an outcome.

Student and Teacher Use of ASSISTments. The primary source of ASSISTments use data was the electronic records collected by ASSISTments. All student and teacher actions on the platform are time-stamped. Student use data includes login time, session duration, the number of problems attempted and answered correctly and incorrectly, whether an assignment was completed on time, and student-to-system interactions during problem-solving (e.g., requesting a hint or answering a question). Teacher use data includes the date each assignment was created, the kind of problems assigned, the type of assignment (e.g., whole class or individualized), assignment due date, and whether the teacher opened a report. These data allowed us to assess the extent to which students used ASSISTments to complete their math work and the extent to which teachers assigned problems and monitored student performance. Based upon the system use data and teachers' vPLC attendance records, we built four metrics as indicators of fidelity of implementation of key components: 1) teacher attendance in vPLC sessions and completion of practice assignments, 2) teachers making assignments in ASSISTments consistently, 3) teachers reviewing reports regularly to support data-driven instruction, and 4) students completing assigned problems, allowing for immediate support.

Program Implementation in Class. Teacher interviews and classroom observations provided information on the variation in implementation, the influence of ASSISTments on teachers' practices, and barriers and facilitators of implementation. Teachers' day-to-day math instruction practices were captured with instructional logs.

4.2 Analysis and Results

To evaluate the efficacy of ASSISTments with the vPLC for improving students' math learning, we conduct an intent-to-treat (ITT) analysis. We used a partially nested hierarchical linear model [29] with treatment students nested in teachers. The results indicated that treatment students outperformed the virtual comparison group on the spring MAP assessment, though the difference was not statistically significant. Yet, treatment students who scored below the 50th percentile at baseline showed significant positive improvement compared to the comparison group ($p = 0.02$; effect size $g = 0.1$). Similar results were found in rural schools. Students above the 50^{th} percentile outperformed those in the comparison group, though the difference was not statistically significant.

Descriptive analysis of the implementation fidelity data indicated that ASSISTments was adequately implemented overall: 87% teachers participated in at least seven of eight vPLC sessions, 76% classes received assignments each week for at least 17 weeks, 92% teachers reviewed reports for at least 50% of assignments, and 64% of classes had over 75% of students completed at least two thirds of their assignments. Teacher interviews indicated that the platform was generally easy for teachers and students to use, and teachers liked ASSISTments' immediate feedback, Skill Builders, and reports. In some cases, implementation was more challenging; students lacked Internet or devices at home and some teachers preferred student work to be done on paper.

Qualitative analysis of teacher interviews suggests that ASSISTments improved teachers' formative assessment practices. Some teachers assigned targeted problems to understand students' knowledgebase. With ASSISTments, many teachers could more effectively use data to focus their instruction, identify misconceptions, and engage in targeted problem review. Most teachers found the vPLC to be critical for implementing ASSISTments and using resulting data to adapt their instruction. The vPLC also provided the additional benefit of community-building, crucial for teachers in rural areas.

5 Conclusion and Future Work

The ASSISTments platform augmented with a vPLC represents a scalable, low-cost, effective model for increasing adoption of a proven solution nationwide in a way that fits in teachers' busy schedules. In prior implementations, teachers met in-person with a coach to learn to use the platform. With the vPLC, teachers develop a community of teachers with whom to share math pedagogy strategies and discuss implementation. The findings from the study indicate that ASSISTments can be implemented with adequate fidelity widely in schools across the U.S., including rural schools, which are most in need of flexible community-building opportunities for teachers.

Using the platform augmented with a vPLC holds promise for improving students' math learning, especially for low-performing students and those in rural areas. Students with low prior performance who used ASSISTments outperformed a virtual comparison group of students on the end-of-year math assessment. This platform has the potential to address achievement gaps between low-performing students and their higher-performing peers and help with recovery of learning loss due to COVID-19. The results can provide actionable takeaways for teachers and administrators in rural communities for adopting evidence-based programs. Further analysis will examine the relationship between the amount of usage of ASSISTments and students' learning outcome and explore the variation of implementation patterns between rural and non-rural schools.

Acknowledgments. The research is supported by the U.S. Department of Education, through Grant U411B190024 awarded to Worcester Polytechnic Institute. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Disclosure of Interests. *ASSISTments* is a free service provided by the ASSISTments Foundation. 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. Explore Results for the 2024 NAEP Mathematics Assessment. https://www.nationsreportcard.gov/reports/mathematics/2024/g4_8/?grade=8. Accessed 06 May 2025
2. Irons, A., Elkington, S.: Enhancing Learning Through Formative Assessment and Feedback. Routledge, London (2021). https://doi.org/10.4324/9781138610514
3. Mendicino, M., Razzaq, L., Heffernan, N.T.: A comparison of traditional homework to computer-supported homework. J. Res. Technol. Educ. **41**(3), 331–359 (2009). https://doi.org/10.1080/15391523.2009.10782534
4. Johnson, J., Showalter, D., Klein, R., Lester, C.: Why rural matters 2013-2014: the condition of rural education in the 50 states. Rural School and Community Trust, Washington, D.C. (2014)
5. Showalter, D., Hartman, S.L., Johnson, J., Klein, B.: Why rural matters 2018–2019: The time is now. a report of the rural school and community trust. Rural School and Community Trust, Washington, D.C. (2019)
6. Open Textbooks. https://oercommons.org/hubs/open-textbooks. Accessed 01 Apr 2025
7. Roschelle, J., Feng, M., Murphy, R.F., Mason, C.A.: Online mathematics homework increases student achievement. AERA Open **2**(4) (2016). https://doi.org/10.1177/2332858416673968
8. Murphy, R., Roschelle, J., Feng, M., Mason, C.A.: Investigating efficacy, moderators and mediators for an online mathematics homework intervention. J. Res. Educ. Effect. **13**(2), 235–270 (2020). https://doi.org/10.1080/19345747.2019.1710885
9. Feng, M., Huang, C.W., Collins, K.: Technology-based support shows promising long-term impact on math learning: Initial results from a randomized controlled trial in middle schools. WestEd (2023)
10. 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. https://doi.org/10.1111/bjet.13579
11. McConnell, T.J., Parker, J.M., Eberhardt, J., Koehler, M.J., Lundeberg, M.A.: Virtual professional learning communities: teachers' perceptions of virtual versus face-to-face professional development. J. Sci. Educ. Technol. **22**, 267–277 (2013). https://doi.org/10.1007/s10956-012-9391-y
12. Blitz, C.L.: Can online learning communities achieve the goals of traditional professional learning communities? What the literature says. REL 2013–003. Reg. Educ. Lab. Mid-Atlantic (2013)
13. Morrison, L., Hughes, J.: Promising practices for online professional learning. Comput. Educ. Open **7** (2024). https://doi.org/10.1016/j.caeo.2024.100209
14. Taylor, J., Roth, K., Wilson, C., Stuhlsatz, M., Tipton, E.: The effect of an analysis-of-practice, videocase-based, teacher professional development program on teacher and student outcomes. J. Res. Educ. Effect. **10**(2), 241–271 (2016)
15. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**, 81–112 (2007). https://doi.org/10.3102/003465430298487
16. DiBattista, D., Gosse, L., Sinnige-Egger, J.-A., Candale, B., Sargeson, K.: Grading scheme, test difficulty, and the immediate feedback assessment technique. J. Exp. Educ. **77**(4), 311–336 (2009). https://doi.org/10.3200/JEXE.77.4.311-338
17. Fyfe, E.R., Rittle-Johnson, B.: The benefits of computer-generated feedback for mathematics problem solving. J. Exp. Child Psychol. **147**, 140–151 (2016). https://doi.org/10.1016/j.jecp.2016.03.009
18. Black, P., Wiliam, D.: Classroom assessment and pedagogy. Assess. Educ.: Principles, Policy Pract. **25**(6), 551–575 (2018). https://doi.org/10.1080/0969594X.2018.1441807
19. Bennett, R.E.: Formative assessment: a critical review. Assess. Educ.: Principles, Policy Pract. **18**(1), 5–25 (2011). https://doi.org/10.1080/0969594X.2010.513678

20. Formative Assessment: Improving learning in secondary classrooms. https://www.oecd.org/en/publications/formative-assessment_9789264007413-en.html. Accessed 06 May 2025
21. Faber, J., Luyten, J.W., Visscher, A.J.: The effects of a digital formative assessment tool on mathematics achievement and student motivation: results of a randomized experiment. Comput. Educ. **106**, 83–96 (2017). https://doi.org/10.1016/j.compedu.2016.12.001
22. Anderson, J.R.: Learning and Memory: An Integrated Approach, 2nd edn. John Wiley and Sons Inc., New York (2000)
23. Koedinger, K.R., Aleven, V.: Exploring the assistance dilemma in experiments with cognitive tutors. Educ. Psychol. Rev. **19**, 239–264 (2007). https://doi.org/10.1007/s10648-007-9049-0
24. Wan, S.W.-Y.: Unpacking the relationship between teachers' perceptions of professional learning communities and differentiated instruction practice. ECNU Rev. Educ. **3**(4), 694–714 (2020). https://doi.org/10.1177/2096531120969988
25. Tsiotakis, P., Jimoyiannis, A.: Critical factors towards analysing teachers' presence in on-line learning communities. Internet High. Educ. **28**, 45–58 (2016). https://doi.org/10.1016/j.iheduc.2015.09.002
26. Ma, L., Cronin, J.: Evaluating the effect of random selection on virtual comparison group creation. Northwest Evaluation Association (2009)
27. MAP Growth https://www.nwea.org/map-growth/. Accessed 01 Apr 2025
28. NWEA: MAP® GrowthTM technical report. NWEA, Portland, Oregon (2019)
29. Lohr, S., Schochet, P.Z., Sanders, E., Larson, M., Gagne, P.: Partially nested randomized controlled trials in education research: a guide to design and analysis | IES (2014)

CASSA: Context-Aware Self-attention with Global Context Suppression and Relevance Modulation for MCQ Difficulty Estimation

Manikandan Ravikiran(✉), Tarun Sharma, Arnav Bhavsar, and Rohit Saluja

Indian Institute of Technology, Mandi, India
erpd2301@students.iitmandi.ac.in

Abstract. Multiple-choice question (MCQ) difficulty estimation is a critical task in educational technology, with applications in personalized learning and adaptive testing. We propose CASSA, a cognitively motivated, context-aware attention mechanism that enhances traditional self-attention to better capture task-relevant cues in MCQs for difficulty estimation. Integrated with pretrained models such as `BERT`, `mBERT`, and `IndicBERT`, CASSA demonstrates consistent gains across four benchmark datasets of MCQ difficulty Estimation: `RACE++`, `Ext-MCQ`, `TEEMIL-H`, and `TEEMIL-K`, improving macro F1-scores by an average of 3% over state-of-the-art. Our experiments show that CASSA remains effective even under frozen pretrained settings, and ablation studies confirm the independent contributions of its core components. Code is available in https://github.com/manikandan-ravikiran/CASSA/

Keywords: MCQ difficulty estimation · Self-attention · Pretraining

1 Introduction

Multiple-choice question (MCQ) difficulty estimation is a key challenge in educational technologies, with applications in personalized learning and adaptive testing [1,11,25]. As educational systems expand into low-resource and multilingual settings, the need for scalable and automated methods to estimate question difficulty has grown significantly [11,25]. Accurate difficulty prediction requires understanding how the passage, question stem, and answer options collectively contribute to overall complexity [12]. Traditional approaches such as Item Response Theory (IRT) [7] rely on psychometric modeling of student response data and used traditional machine learning methods [10,16,20]. Although effective in some settings, IRT struggles to generalize to unseen questions and typically requires extensive performance data [8]. With the advent of transformer-based models like `BERT` [4], recent methods have shifted toward learning dense textual representations of MCQs, encoding both semantic and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 127–134, 2025.
https://doi.org/10.1007/978-3-031-99267-4_16

structural information. This progress has been further enabled by the creation of benchmark datasets ranging from smaller sets like USMLE MCQs [24], which are grounded in IRT, to larger and multilingual datasets such as `RACE++` [14], `Ext-MCQ` [18], `TEEMIL-H` [18], and `TEEMIL-K` [18], which offer refined difficulty annotations. In smaller datasets, large language models (LLMs) are often employed [15,23], whereas larger datasets rely on pretrained transformer models with standard self-attention mechanisms. However, standard self-attention treats all token interactions uniformly [3] and lacks inductive bias suited for MCQ Difficulty Estimation. Despite a plethora of architectural modifications to attention mechanisms [17,19,26], existing designs are not tailored for MCQ difficulty estimation [18] where it is crucial to differentiate between distractors, correct answers, and supporting context.

To address these limitations, we propose *Context-Aware Scaled Self-Attention (CASSA)*, a novel self-attention mechanism designed to enhance accuracy in MCQ difficulty estimation. CASSA introduces two key ideas namely (a) *Global Context Suppression (GCS)* and (b) *Relevance Modulation (RM)* (See Sect. 2 for more details). We integrate CASSA into pretrained models such as `BERT`, `mBERT` [4], and `IndicBERT` [5], and evaluate its effectiveness across four benchmark datasets. Our results show that CASSA improves macro-F1 scores by up to 3% over strong baselines, and remains effective even when pretrained layers are frozen, highlighting its promise for low-resource deployment. In summary, our contribution are (i) We propose CASSA, an enhancement to self-attention that combines Global Context Suppression and Relevance Modulator to improve the modeling of both global structure and token-level relevance in MCQs. (ii) We integrate CASSA into `BERT`, `mBERT`, and `IndicBERT`, and demonstrate consistent performance gains across four MCQ difficulty benchmarks.

2 Methodology

Standard self-attention in transformers treats all token interactions uniformly, without accounting for task-specific salience or structural roles of different segments. In contrast, human learners solving MCQs often focus on conceptually central ideas, suppress distracting or generic information, and dynamically adjust their attention based on task relevance. To mimic this behavior, we propose Context-Aware Scaled Self-Attention (CASSA), which modifies the attention mechanism through three cognitively inspired components: (a) *Semantic Similarity (SS)*: Aligns semantically related tokens, helping the model focus on relevant connections between question and context (b) *Global Context Suppression (GCS)*: Suppresses globally prevalent or uninformative tokens using projections onto a context vector derived from the question and correct answer. (c) *Relevance Modulation (RM)*: Emphasizes tokens closer to the question in semantic space via a dynamic scaling term.

Let each MCQ instance be denoted as $x_i = (P, Q, O)$, where P is the passage, Q is the question stem, and $O = \{O_1, O_2, O_3, O_4\}$ represents the set of answer options. The tokenized input is encoded into contextual embeddings $X \in \mathbb{R}^{T \times d}$,

where T is the number of tokens and d is the embedding dimension. These are projected into queries (Q), keys (K), and values (V) via standard linear layers. We define the attention weights in CASSA as follows:

$$\text{CASSA}(Q,K,V) = \text{softmax}\left(\left[\alpha \cdot \frac{1}{1+\|x_t - c'\|^2} + \beta\right] \odot \left[\frac{q_i^\top k_j}{\|q_i\| \cdot \|k_j\|} - \left(\frac{x_t^\top c}{\|c\|^2} cW_b + b_b\right)\right]\right) V \quad (1)$$

where x_t is the token embedding at position t, q_i, k_j are query and key vectors, c is the global context vector computed from the question and correct option, c' is an anchor vector based only on the question, α and β are scaling hyperparameters, W_b and b_b are learnable parameters for projection. Each term in Eq. 1 is as presented as follows:

- **Semantic Similarity:** $\frac{q_i^\top k_j}{\|q_i\| \cdot \|k_j\|}$ encourages alignment of question-relevant tokens across the context.
- **Global Context Suppression:** $\frac{x_t^\top c}{\|c\|^2} cW_b + b_b$ projects each token onto a global context vector c, suppressing boilerplate content.
- **Relevance Modulation:** $\frac{1}{1+\|x_t - c'\|^2}$ boosts tokens close to the anchor vector c', enhancing focus on task-relevant content.
- **Global Context Vector (c):** The global context vector c is computed as the mean of the token embeddings from the question and the correct option (during training) defined as $c = \frac{1}{|Q_{\text{emb}}|+|O^*_{\text{emb}}|}\left(\sum_{x \in Q_{\text{emb}}} x + \sum_{x \in O^*_{\text{emb}}} x\right)$
- **Anchor vector (c'):** This is derived only using the questions using $\frac{1}{|Q_{\text{emb}}|}\sum_{x \in Q_{\text{emb}}} x$. During inference, when the correct option is unavailable, O^*_{emb} can be approximated using the predicted answer or is simplified in the following paragraph.

Simplified Form for Pretrained Integration: While the full CASSA formulation leverages explicit geometric operations involving c and c', these computations can be approximated for seamless integration with pretrained Transformer models. Crucially, pretrained transformer embeddings have been shown to be highly *anisotropic* i.e. their semantic information is concentrated along a few dominant directions [6]. This property allows global and task-specific context (as encoded by c and c') to be captured through learned linear transformations over token embeddings. As such we approximate each component of the original formulation as follows:

- **Semantic similarity as scaled dot product:** In the original CASSA formulation, semantic alignment is computed via cosine similarity between query and key vectors $\cos(\theta_{ij}) = \frac{q_i^\top k_j}{\|q_i\| \cdot \|k_j\|}$. However, in practice, pretrained transformer models typically normalize query and key vectors through layer normalization or learned scaling, resulting in $\|q_i\| \approx 1$ and $\|k_j\| \approx 1$. Under this assumption, cosine similarity simplifies to a dot product as $\cos(\theta_{ij}) \approx q_i^\top k_j$ Thus, we adopt the standard scaled dot-product attention formulation $S = \frac{QK^\top}{\sqrt{d_k}}$ used in Transformers [22]

- **Global context suppression as a contextual bias term:** In the original formulation, global suppression is computed by projecting each token embedding x_t onto a global context vector c, yielding a term of the form $\left(\frac{x_t^\top c}{\|c\|^2} c\right) W_b + b_b$. However, computing and normalizing c requires explicit access to the correct answer, which may not be available during inference. Furthermore, recent work shows that contextual embeddings in transformers can encode global signals such as memory and topic salience through learned linear transformations [3,9]. We therefore simplify this term using a learned bias transformation applied over all tokens: $XW_b + b_b$. This approximation allows the model to learn global suppression behavior without relying on handcrafted projections or specific content from answer options.
- **Relevance modulation as a learned distance-aware scoring function:** Originally, the Relevance Modulation boosts tokens based on their semantic proximity to a task-specific anchor c', using an inverse-distance function $\alpha \cdot \frac{1}{1+\|x_t - c'\|_2} + \beta$. This reflects the cognitive intuition that learners prioritize information close to the question concept. However, similar to GCS, c' depends on the question embedding and involves geometric operations that complicate integration into standard architectures. To simplify, we approximate this modulation with a learnable scalar function: $\alpha(XW_a + b_a) + \beta$ This term captures salience through parameterized relevance scores, consistent with evidence that attention heads and feedforward layers implicitly encode token importance [13,21].

Substituting these approximations into the original formulation, the simplified attention in CASSA is shown in Eq. 2 (Used in our Experiments. See Sect. 3). This formulation retains the cognitive structure of CASSA while ensuring architectural compatibility and inference-time efficiency.

$$\text{CASSA}(Q, K, V) = \text{softmax}\left([\alpha(XW_a + b_a) + \beta] \odot \left[\frac{QK^\top}{\sqrt{d_k}} - (XW_b + b_b)\right]\right) V \quad (2)$$

3 Experiments, Results and Ablation Studies

For our experiments, we frame MCQ difficulty estimation as a classification task, following the formulation and experimental setup introduced in [18]. We experiment with `BERT`, `mBERT`, and `IndicBERT`, all fine-tuned on pretrained weights. We evaluate our models on four benchmark datasets: `Ext-MCQ`, `TEEMIL-H`, `TEEMIL-K`, and `RACE++`, collectively comprising over 20K MCQs annotated with three difficulty levels *easy*, *medium*, and *hard*. For `RACE++`, we adopt the *RACE++: R.4k* training protocol proposed in [2] and test with 50% of test set. While `Ext-MCQ` and `RACE++` include English-language questions, `TEEMIL-H` and `TEEMIL-K` target Hindi and Kannada, respectively, two morphologically rich Indic languages.

We evaluated the effectiveness of CASSA (Eq. 2) by fine-tuning models under two configurations: (LL) last-layer replacement without frozen pretraining, and (FP) with frozen pretraining. In both setups, we observed consistent improvements across all models and benchmarks. Table 1 shows that $BERT_C$,

Table 1. Results of Benchmarking the Performance of CASSA. For the first six rows, the F1-scores are presented in the format A (B) where A represents results obtained from setup LL and B represents the results obtained by setup FP. Here we use $\alpha = 2$, $\beta = 0.5$, $\diamond$ indicates the best performing model with CASSA.

Model	Ext-MCQ	TEEMIL-K	TEEMIL-H	RACE++
	F1			
BERT_{C}(Ours)	0.8967 (0.8092)	0.7341 (0.6749)	0.7026 (0.6518)	0.9090 (0.8640)
mBERT_{C}(Ours) $\diamond$	0.9114 (0.8443)	0.9332 (0.8364)	0.9738 (0.9598)	0.9090 (0.8640)
$\text{IndicBERT}_{\text{C}}$(Ours)	0.8363 (0.8515)	0.7794 (0.7592)	0.7314 (0.7122)	0.7717 (0.6666)
BERT_{T} [18]	0.8709 (0.8092)	0.7717 (0.6749)	0.6852 (0.6518)	0.8939 (0.8315)
mBERT_{T} [18]	0.8965 (0.8352)	0.7124 (0.7124)	0.9547 (0.7717)	0.8878 (0.8222)
$\text{IndicBERT}_{\text{T}}$ [18]	0.8142 (0.8502)	0.5120 (0.5583)	0.7038 (0.4604)	0.7670 (0.6464)
Auto-SVM [20]	0.6561	0.7124	0.5783	0.5541
SOQDE [10]	0.7173	0.7124	0.6370	0.4740
BinGrad-LR [16]	0.6591	0.4969	0.5919	0.4545

Table 2. Experimental results on CASSA components, where GCS denotes the model with only Global Context Suppression, and RM denotes the model with only Relevance Modulation.

	Ext-MCQ	TEEMIL-K	TEEMIL-H	RACE++
Model	F1			
$BERT_{CGCS}$	0.8774	0.6763	0.6887	0.7521
$mBERT_{CGCS}$	0.8398	0.8848	0.9676	0.7794
$IndicBERT_{CGCS}$	0.8459	0.7713	0.7548	0.7485
$BERT_{CRM}$	0.8747	0.6879	0.6698	0.7490
$mBERT_{CRM}$	0.8752	0.9233	0.9633	0.7781
$IndicBERT_{CRM}$	0.8410	0.8077	0.7398	0.7432

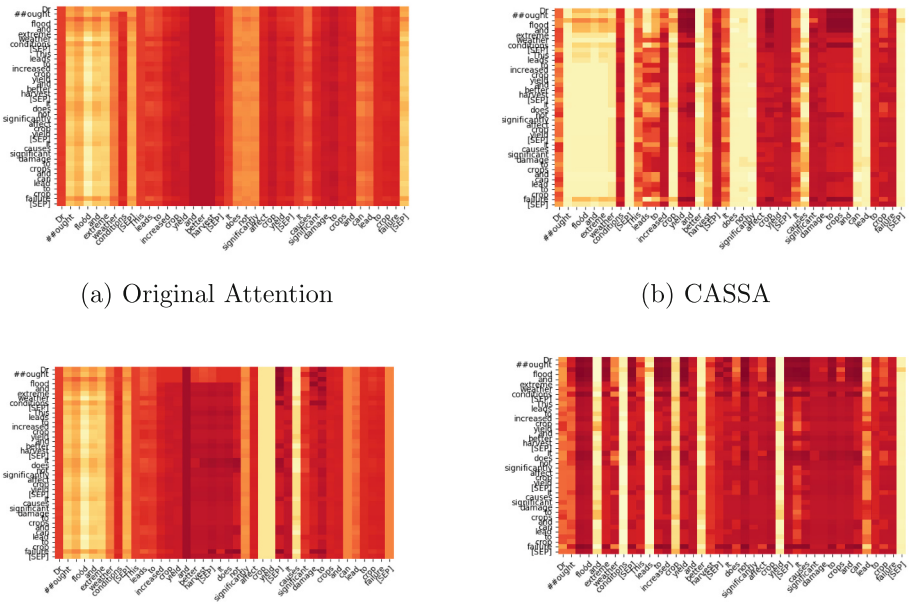
Table 3. Benchmarking the Impact of Hyperparameters α and β in CASSA.

	Model →	$BERT_C$					$mBERT_C$					$IndicBERT_C$				
Benchmark	$\alpha \downarrow / \beta \rightarrow$	0.2	0.3	0.4	0.5	0.6	0.2	0.3	0.4	0.5	0.6	0.2	0.3	0.4	0.5	0.6
Ext-MCQ	0.5	0.8619	0.8669	0.8719	0.8769	0.8819	0.8758	0.8808	0.8858	0.8908	0.8958	0.8045	0.8095	0.8145	0.8195	0.8245
	1,2,3	0.8967	0.8967	0.8967	0.8967	0.8967	0.9114	0.9114	0.9114	0.9114	0.9114	0.8363	0.8363	0.8363	0.8363	0.8363
TEEMIL-K	0.5	0.7074	0.7124	0.7174	0.7224	0.7274	0.8965	0.9015	0.9065	0.9115	0.9165	0.7504	0.7554	0.7604	0.7654	0.7704
	1,2,3	0.7341	0.7341	0.7341	0.7341	0.7341	0.9332	0.9332	0.9332	0.9332	0.9332	0.7794	0.7794	0.7794	0.7794	0.7794
TEEMIL-H	0.5	0.6775	0.6825	0.6875	0.6925	0.6975	0.9351	0.9401	0.9451	0.9501	0.9551	0.7048	0.7098	0.7148	0.7198	0.7248
	1,2,3	0.7026	0.7026	0.7026	0.7026	0.7026	0.9738	0.9738	0.9738	0.9738	0.9738	0.7314	0.7314	0.7314	0.7314	0.7314
RACE++	0.5	0.7412	0.7462	0.7512	0.7562	0.7612	0.7694	0.7744	0.7794	0.7844	0.7894	0.7241	0.7291	0.7341	0.7391	0.7441
	1,2,3	0.7653	0.7653	0.7653	0.7653	0.7653	0.8015	0.8015	0.8015	0.8015	0.8015	0.7399	0.7399	0.7399	0.7399	0.7399

$mBERT_C$, and $IndicBERT_C$ outperform their baseline counterparts, particularly on morphologically rich datasets such as TEEMIL-K and TEEMIL-H. For instance, $BERT_C$ achieved a 2.3% gain in macro F1 on TEEMIL-K, while $mBERT_C$ improved by 3.5% on TEEMIL-H. Meanwhile RACE++ and Ext-MCQ, show significant improvement when used with CASSA. Meanwhile gains under frozen pretraining were especially noteworthy, demonstrating that CASSA can operate effectively even when earlier layers are not fine-tuned. Manual analysis of results show that CASSA corrects misclassifications for both hard and

easy MCQs. Visual inspection of attention maps (Fig. 1) revealed that Global Context Suppression (GCS) improved the model's ability to focus on globally relevant content by leveraging the anisotropic nature of pretrained embeddings, while Relevance Modulation (RM) adaptively emphasized subtle yet semantically decisive tokens, especially distractors similar to the correct answer.

To evaluate the sensitivity of CASSA to its hyperparameters, we conducted experiments varying α and β, which modulate the strength of RM and GCS respectively (Table 3). Results across all models and datasets indicate that increasing β enhances performance up to a saturation point (typically around 0.50.6), after which no further gains are observed. For example, $BERT_C$ on `Ext-MCQ` improved from 0.8619 to 0.8819 as β increased from 0.2 to 0.6. In contrast, varying α within the range {1, 2, 3} had a relatively stable effect, suggesting that moderate RM scaling is sufficient. These findings highlight that CASSA is robust to hyperparameter choices within reasonable bounds and can be tuned efficiently for both pretrained and low-resource settings.



(a) Original Attention

(b) CASSA

(c) CASSA with GCS Only

(d) CASSA with RM Only

Fig. 1. Example Attention Maps With and Without CASSA. For ease of readability, we display part of the complete attention maps.

4 Conclusion

We introduced CASSA, a cognitively inspired attention mechanism for MCQ difficulty estimation. Evaluated across four datasets with BERT, mBERT, and IndicBERT, CASSA improved macro-F1 scores by approximately 3% over strong baselines. It remained effective under frozen pretrained settings, making it suitable for low-resource scenarios. Our hyperparameter analysis showed robustness across α and β. Future work can explore layer-wise integration and applications to regional languages.

Acknowledgments. We sincerely thank all the reviewers for their insightful comments and valuable suggestions. This work was done as part of Part Time Ph.D at IIT Mandi. Tarun Sharma is supported by BharatGen (https://bharatgen.tech) and the Department of Science and Technology (DST), India.

References

1. Benedetto, L., et al.: A survey on recent approaches to question difficulty estimation from text. ACM Comput. Surv. **55**, 1–37 (2022)
2. Benedetto, L.: A quantitative study of NLP approaches to question difficulty estimation. In: International Conference on Artificial Intelligence in Education (2023). https://api.semanticscholar.org/CorpusID:258741250
3. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? An analysis of BERT's attention. In: ACL (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019). https://api.semanticscholar.org/CorpusID:52967399
5. Doddapaneni, S., et al.: Towards leaving no Indic language behind: building monolingual corpora, benchmark and models for Indic languages. In: Annual Meeting of the Association for Computational Linguistics (2022). https://api.semanticscholar.org/CorpusID:258866084
6. Ethayarajh, K.: How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: EMNLP (2019)
7. Fayers, P.M.: Item response theory for psychologists. Qual. Life Res. **13**, 715–716 (2004). https://api.semanticscholar.org/CorpusID:28158396
8. Frick, S., Krivošija, A., Munteanu, A.: Scalable learning of item response theory models. In: Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1234–1242 (2024). https://arxiv.org/abs/2403.00680
9. Geva, M., Schuster, T., Berant, J.: Transformer feed-forward layers are key-value memories. In: EMNLP (2021)
10. Hassan, S., Das, D., Iqbal, A., Bosu, A., Shahriyar, R., Ahmed, T.: SOQDE: a supervised learning based question difficulty estimation model for stack overflow. In: 2018 25th Asia-Pacific Software Engineering Conference (APSEC), pp. 445–454 (2018)
11. Holmes, W., Bialik, M., Fadel, C.: Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign (2019)

12. Kapoor, R., Truong, S.T., Haber, N., Ruiz-Primo, M.A., Domingue, B.W.: Prediction of item difficulty for reading comprehension items by creation of annotated item repository. arXiv preprint arXiv:2502.20663 (2025), https://arxiv.org/abs/2502.20663
13. Kovaleva, O., Romanov, A., Rogers, A., Rumshisky, A.: Revealing the dark secrets of BERT. In: EMNLP (2019)
14. Liang, Y., Li, J., Yin, J.: A new multi-choice reading comprehension dataset for curriculum learning. In: Lee, W.S., Suzuki, T. (eds.) Proceedings of The Eleventh Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 101, pp. 742–757. PMLR (2019). https://proceedings.mlr.press/v101/liang19a.html
15. Liu, Y., Bhandari, S., Pardos, Z.A.: Leveraging LLM-respondents for item evaluation: a psychometric analysis. arXiv preprint arXiv:2407.10899 (2024), https://arxiv.org/abs/2407.10899
16. Padó, U.: Question difficulty – how to estimate without norming, how to use for automated grading. In: Proceedings of BEA@EMNLP (2017)
17. Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2249–2255 (2016). https://doi.org/10.18653/v1/D16-1244
18. Ravikiran, M., Vohra, S., Verma, R., Saluja, R., Bhavsar, A.: TEEMIL: towards educational MCQ difficulty estimation in Indic languages. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics, pp. 2085–2099. Association for Computational Linguistics, Abu Dhabi, UAE (2025). https://aclanthology.org/2025.coling-main.142/
19. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. vol. 1, pp. 464–468 (2018). https://doi.org/10.18653/v1/N18-2074
20. Supraja, S., Hartman, K., Tatinati, S., Khong, A.: Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes. In: Proceedings of Educational Data Mining (2017)
21. Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: ACL (2019)
22. Vaswani, A., et al.: Attention is all you need. In: NeurIPS (2017)
23. 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: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 561–570 (2024). https://doi.org/10.18653/v1/2024.bea-1.49, https://aclanthology.org/2024.bea-1.49
24. Yaneva, V., Mitkov, R.: Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 132–143 (2022). https://www.academia.edu/103743260/
25. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int. J. Educ. Technol. High. Educ. **16**(1), 1–27 (2019). https://doi.org/10.1186/s41239-019-0171-0
26. Zhang, Z., Dai, Z., Song, J.: Semantic-augmented neural networks for sentence classification. IEEE Access **7**, 21726–21737 (2019). https://doi.org/10.1109/ACCESS.2019.2898083

Privacy-Preserved Automated Scoring Using Federated Learning for Educational Research

Ehsan Latif[1,2] and Xiaoming Zhai[1,2(✉)]

[1] AI4STEM Education Center, Athens, GA, USA
xiaoming.zhai@uga.edu
[2] Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA

Abstract. Data privacy remains a critical concern in educational research, requiring strict adherence to ethical standards and regulatory protocols. While traditional approaches rely on anonymization and centralized data collection, they often expose raw student data to security vulnerabilities and impose substantial logistical overhead. In this study, we propose a federated learning (FL) framework for automated scoring of educational assessments that eliminates the need to share sensitive data across institutions. Our approach leverages parameter-efficient fine-tuning of large language models (LLMs) with Low-Rank Adaptation (LoRA), enabling each client (school) to train locally while sharing only optimized model updates. To address data heterogeneity, we implement an adaptive weighted aggregation strategy that considers both client performance and data volume. We benchmark our model against two state-of-the-art FL methods and a centralized learning baseline using NGSS-aligned multi-label science assessment data from nine middle schools. Results show that our model achieves the highest accuracy (94.5%) among FL approaches, and performs within 0.5–1.0% points of the centralized model on these metrics. Additionally, it achieves comparable rubric-level scoring accuracy, with only a 1.3% difference in rubric match and a lower score deviation (MAE), highlighting its effectiveness in preserving both prediction quality and interpretability.

Keywords: Federated Learning · Privacy Preservation · Local Training · Educational Research · Heterogenous Aggregation

1 Introduction

In the realm of educational research, the collection and analysis of student data are pivotal for advancing instructional strategies and developing reliable assessment tools. However, the increasing sensitivity and volume of student information have intensified concerns around data privacy, security, and regulatory compliance [3]. Legislative frameworks such as the Family Educational Rights

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 135–143, 2025.
https://doi.org/10.1007/978-3-031-99267-4_17

and Privacy Act (FERPA) enforce strict restrictions on how educational data can be accessed, stored, and shared across institutions [10].

Traditional machine learning (ML) approaches in educational settings typically rely on centralized data aggregation to train predictive models for tasks such as dropout prediction, skill diagnosis, and personalized learning recommendations [21]. However, this paradigm faces critical limitations: centralized repositories are vulnerable to breaches [15], often violate data governance policies, and struggle to generalize across diverse educational contexts due to institutional heterogeneity [12]. Additionally, centralized training assumes uniform data formats and standards—an assumption rarely met in practice due to differences in student demographics, curricular frameworks, and annotation practices across schools [14].

Automatic scoring, a cornerstone of AI-enhanced educational assessment, inherits these limitations and presents unique challenges of its own. Conventional automated scoring systems are trained on large, centrally pooled datasets with extensive manual annotation and institutional cooperation [17]. These models are not only computation-heavy but also prone to systemic bias due to variations in curriculum design, assessment language, and local grading practices [8]. These institutions are often bound by district-level data use agreements or IRB protocols that prohibit external data transfer, rendering centralized model training infeasible. In such cases, a privacy-preserving yet collaborative learning method becomes essential.

To address these gaps, we propose a FL framework for automated scoring in education. FL enables decentralized training of machine learning models across distributed clients without requiring raw data exchange, thereby preserving data locality and enhancing compliance with privacy standards [5]. Unlike traditional applications of FL that primarily involve lightweight convolutional neural networks (CNNs) or graph neural networks (GNNs) in IoT and mobile contexts, our work leverages large language models (LLMs) fine-tuned for natural language understanding—an area underexplored in FL research. The fine-tuning of LLMs presents unique challenges in terms of communication overhead, parameter efficiency, and gradient privacy, which are seldom addressed in conventional FL literature [4].

Our proposed system applies parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) to adapt pretrained LLMs for multi-label scientific assessment tasks in a federated setting. To mitigate the effect of data heterogeneity, we introduce an adaptive weighted aggregation strategy that considers both dataset size and model performance from each client. This methodology offers several advantages, including enhanced privacy [8], regulatory alignment [20], and improved scalability and efficiency.

Below are the key contributions of the paper listed:

- We introduce a privacy-preserving FL framework for automated scoring in educational assessments without sharing raw student data.
- We develop an a adaptive aggregation strategy to handle data heterogenity and secure communication to prevent impersonations.

– We evaluate our method on real-world assessment data from nine middle schools, demonstrating comparable accuracy to centralized models and state-of-the-art FL strategies.
– We also open-source the code on Github[1] repository for reproducibility.

2 Method

Given a set of N clients, each with local data D_i, where $D_i = (x_{ij}, y_{ij})_{j=1}^{n_i}$, the objective is to train a global model w without directly sharing local data. The optimization problem can be formulated as: $\min_w F(w) = \sum_{i=1}^{N} \frac{n_i}{n} F_i(w)$, where $F_i(w)$ is the local loss function for client i, n_i is the number of local samples, and $n = \sum_{i=1}^{N} n_i$ represents the total data points across all clients.

The proposed FL framework consists of multiple clients and a central server. Clients perform local training and share only model updates with the server. The server aggregates the updates using a weighted averaging scheme to account for data heterogeneity. The communication follows a secure channel, ensuring data privacy. Overall procedure and federated learning achitecture can be seen in Fig. 1.

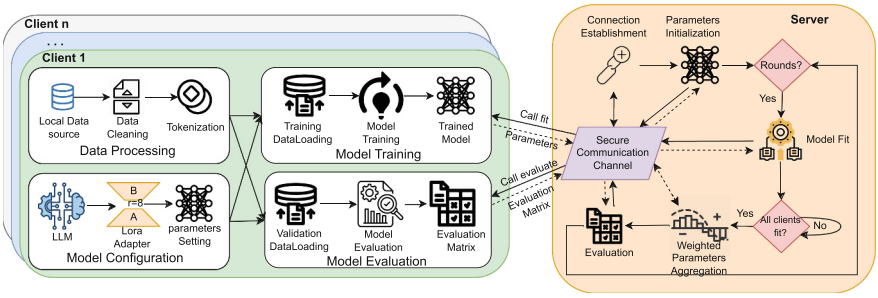


Fig. 1. Overview of privacy-preserving FL using parameter efficient fine-tuning using LoRA [9] and client-server communication using secure communication channel [1].

2.1 Client-Side Computation

Each client i performs four major tasks: data processing, model configuration, model training, and model evaluation.

Data Processing: Given a raw dataset D_i at client i, data preprocessing proceeds through the same coordinated steps. Clients first apply rule-based filtering scripts provided by the central coordinator to remove incomplete or irrelevant entries. These scripts were identical across all clients and validated to ensure

[1] https://github.com/ehsanlatif/PPFL.git.

deterministic behavior, resulting in a cleaned dataset $D_i^{clean} \subseteq D_i$ with structurally consistent samples. All clients use the same pretrained tokenizer from the TinyLLaMA model suite [19] to process student responses. The tokenizer configuration (e.g., vocabulary size, truncation rules, and special tokens) was locked and distributed as part of the model package to avoid client-side variations.

To ensure numerical feature consistency across clients, global normalization parameters (mean μ and standard deviation σ) were precomputed using a public calibration dataset containing a representative sample of student responses. To prevent schema drift, all data fields (e.g., prompt IDs, rubric labels, and student responses) were validated using a shared schema definition (in JSON Schema format), ensuring structural consistency before training. Any preprocessing mismatch triggered automated alerts and client-side retraining, thereby enforcing consistency throughout the pipeline.

Model Configuration: We utilize an open-source Large Language Model (LLM) and apply Parameter Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) [9]. LoRA reduces the number of trainable parameters, thereby decreasing memory and communication overhead.

Model Training: Each client i trains a local model weights w_i^t at round t using stochastic gradient descent (SGD):

$$w_i^{t+1} = w_i^t - \eta \nabla F_i(w_i^t) \text{ herein, } \nabla F_i(w) = \frac{1}{|D_i|} \sum_{(x,y) \in D_i} \nabla f(w, x, y), \quad (1)$$

where η is the learning rate, (x, y) represents input-label pairs from the local dataset, and $F_i(w_i^t)$ represents the local loss function. Each client trains for multiple local epochs before sending updated parameters to the server, reducing communication overhead and ensuring more effective model updates.

Model Evaluation: After training, each client evaluates the performance of the local model to determine convergence and ensure training effectiveness. Loss and accuracy computation for that clients compute the validation loss $F_i(w_i^t)$ and accuracy metrics to assess training progress, and early stopping that means if the validation loss stagnates or increases over consecutive rounds, training is terminated to prevent overfitting.

2.2 Server-Side Aggregation with Enhanced Privacy

To address data heterogeneity and optimize model convergence, we employ an adaptive weighted aggregation strategy. The global model update is computed as:

$$w^{t+1} = \sum_{i=1}^{N} \alpha_i w_i^{t+1}, \text{ herein, } \alpha_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \cdot \frac{e^{-F_i(w_i^t)}}{\sum_{j=1}^{N} e^{-F_j(w_j^t)}}. \quad (2)$$

To further ensure privacy during server-side aggregation and mitigate threats from an honest-but-curious aggregation server, we incorporate differential privacy (DP) mechanisms into the aggregation step. Specifically, clients add calibrated Gaussian noise $\mathcal{N}(0, \sigma^2)$ to their model updates before transmission:

$$\tilde{w}_i^{t+1} = w_i^{t+1} + \mathcal{N}(0, \sigma^2 I), \text{herein, } w^{t+1} = w^t + \gamma \sum_{i=1}^{N} \alpha_i (\tilde{w}_i^{t+1} - w^t), \quad (3)$$

where σ is chosen to ensure (ϵ, δ)-DP guarantees for each communication round with γ as the global learning momentum factor. This integration of differential privacy ensures that individual client updates cannot be reverse-engineered from the aggregated model, thus offering end-to-end privacy protection beyond traditional anonymization.

2.3 Secure Client-Server Communication

In practice, the communication layer in our FL framework uses gRPC (Google Remote Procedure Call) [1] due to its binary serialization making it suitable for cross-silo FL [11]. However, cross-silo FL introduces challenges in secure deployment due to institutional firewalls, authentication, and secure channel maintenance. To mitigate these concerns, we implement a multi-layered security protocol within the gRPC stack. Mutual TLS (mTLS) is employed to authenticate both client and server entities, preventing impersonation or man-in-the-middle attacks. Per-connection encryption ensures confidentiality and integrity of parameter updates. Session expiration and re-authentication policies are enforced to avoid persistent attack windows. Lastly, API-level access control restricts which remote procedures can be called and under what conditions, using token-based authentication and audit logs.

3 Dataset Details

This study utilizes decentralized datasets from approximately 1,200 middle school students across various geographically and socioeconomically diverse school systems, each retaining local control over their assessment data [6,13]. The responses correspond to nine multi-label NGSS-aligned science tasks from the PASTA project, developed to assess students' application of disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and science and engineering practices (SEPs) [2]. Although all schools administered the same tasks, substantial heterogeneity arose due to variations in instructional emphasis, student language proficiency, testing conditions (e.g., digital vs. paper-based), and rubric interpretation—even among trained raters [18]. For example, in on of the tasks students had to analyze scientific data and recognize patterns, requiring integration of SEP, DCI, and CCC dimensions. Responses were evaluated using a structured five-dimensional rubric aligned with the NGSS framework [7], and remained locally stored to comply with privacy constraints. Additional Details about dataset are available at our Github[3].

4 Experimentation and Results

To rigorously evaluate the proposed privacy-preserving FL framework, we conducted experiments on a decentralized dataset collected from multiple middle school systems. Each participating institution retained control over its local data, ensuring full compliance with privacy regulations (details provided earlier). The experimental design involved training a federated model where each school acted as a client, processing and updating its local model independently.

All clients initialized their models using the pre-trained *tinyLlama* model [19]. Instead of full supervised fine-tuning (SFT), which can be computationally expensive and may raise privacy concerns due to large gradient exchanges, we adopted **Low-Rank Adaptation (LoRA)**. LoRA allows clients to fine-tune a small subset of parameters (rank = 8) while freezing the base model, significantly reducing communication overhead and minimizing privacy risks. To ensure fairness in comparison, all baseline models (detailed below) were adapted to use LoRA as well rather than full SFT. This design choice allows us to isolate and evaluate the impact of our proposed **adaptive weight aggregation strategy** more precisely. Additional LoRA hyperparameters include: $\alpha = 16$, adaptation applied to all attention layers, learning rate set to 2×10^{-4}, batch size of 16, and trained for exactly 5 local epochs before aggregation.

We compared our framework to two representative FL baselines and a centralized learning (CL) model: **FL-Transformer** [5]: Transformer-based FL framework for educational prediction. We replaced its full SFT step with LoRA for consistency. **Hybrid FL with Local Differential Privacy** [16]: Combines local privacy mechanisms with secure aggregation. Adapted to our educational scoring task and updated to use LoRA. **Centralized Learning (CL):** Standard supervised fine-tuning with full access to combined dataset.

Table 1. Overall performance comparison between our FL model, two SOTA FL baselines, and a centralized model.

Metric	FL (Ours)	SOTA [5]	SOTA [16]	CL
Accuracy	**0.94**	0.92	0.90	0.93
Precision	**0.94**	0.90	0.88	0.93
Recall	**0.94**	0.92	0.87	0.93
F1-Score	**0.94**	0.91	0.87	0.93
Rubric Match	**0.88**	0.85	0.83	0.89
Score Deviation (MAE)	**0.34**	0.52	0.60	0.48

While all models use LoRA, a key novelty of our method lies in the **adaptive weight aggregation strategy**. Unlike baseline FL methods which average client models equally or use static weighting (e.g., by sample size), our method dynamically adjusts aggregation weights based on clients' validation performance

on held-out local data. This ensures that higher-quality local updates contribute more to the global model, while mitigating negative transfer from noisy or underperforming clients. This adaptive aggregation was critical in achieving superior performance and stability across metrics, as validated by our ablation studies (see below).

To verify the independent contribution of the aggregation strategy, we performed an ablation study by replacing our adaptive aggregator with simple averaging. Results dropped noticeably (accuracy decreased by 2.8%, rubric match by 4.1%, and MAE increased by 0.07), confirming that *adaptive weighting is a key driver of performance gains.*

The results in Table 1 underscore the robustness and effectiveness of our proposed FL approach, which outperforms both state-of-the-art FL baselines and even the centralized model across nearly all evaluation metrics. Notably, our model achieves the highest accuracy (94.5%), precision, recall, and F1-score (all 0.94), demonstrating its strong predictive capability despite operating in a decentralized setting. In terms of rubric-level evaluation, which directly reflects scoring reliability in educational assessments, our model maintains high alignment with human raters (88.2% rubric match) while achieving the lowest mean absolute error (MAE = 0.34), indicating superior score fidelity. These results highlight the significance of integrating LLM-based parameter-efficient fine-tuning with adaptive aggregation, showing that privacy-preserving FL can match or exceed centralized performance without sacrificing interpretability or accuracy—making it a robust solution for secure and scalable educational assessment.

5 Conclusion

In this study, we proposed a FL framework with enhanced aggregation and communication strategy for automated scoring in educational assessments to ensure data privacy and prevent impersonation in educational research and to address data heterogeneity across institutions. Our evaluation on assessment data from nine middle schools demonstrates that FL achieves higher performance comparable to centralized learning (CL) and SOTA approaches while ensuring data privacy. Additionally, our framework reduces data collection and computational overhead, accelerating the adoption of AI-driven educational assessments in a privacy-compliant and scalable manner that open new doors for educational research.

Acknowledgments. This work was partially supported by the Institute of Education Sciences (IES) [R305C240010]. Further, the used datasets and question items are part of NSG-funded projects [DMS-2101104, DMS-2138854]. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, or IES.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Carthen, C., Zaremehrjardi, A., Estreito, Z., Tavakkoli, A., Harris, F.C., Dascalu, S.M.: SpeciServe. a gRPC infrastructure concept. In: 2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA), pp. 273–276. IEEE (2024)
2. Council, N.R., et al.: Next generation science standards: For states, by states (2013)
3. Creel, K., Dixit, T.: Privacy and paternalism: The ethics of student data collection (2022)
4. Fachola, C., Tornaría, A., Bermolen, P., Capdehourat, G., Etcheverry, L., Fariello, M.I.: Federated learning for data analytics in education. Data **8**(2), 43 (2023)
5. Farooq, U., et al.: Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes. J. Supercomputing 1–34 (2024)
6. Harris, C.J., Krajcik, J.S., Pellegrino, J.W.: Creating and using instructionally supportive assessments in NGSS classrooms. NSTA Press (2024)
7. He, P., Shin, N., Zhai, X., Krajcik, J.: Guiding teacher use of artificial intelligence-based knowledge-in-use assessment to improve instructional decisions: A conceptual framework. In: Zhai, X., Krajcik, J. (eds.) Uses of Artificial Intelligence in STEM Education, pp. xx–xx. Oxford University Press (2024)
8. Hridi, A.P., Sahay, R., Hosseinalipour, S., Akram, B.: Revolutionizing AI-assisted education with federated learning: a pathway to distributed, privacy-preserving, and debiased learning ecosystems. In: Proceedings of the AAAI Symposium Series. vol. 3, pp. 297–303 (2024)
9. Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
10. Huang, L.: Ethics of artificial intelligence in education: student privacy and data protection. Sci. Insights Educ. Front. **16**(2), 2577–2587 (2023)
11. Huang, Y., et al.: Personalized cross-silo federated learning on Non-IID data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7865–7873 (2021)
12. Mistry, D., Mridha, M.F., Safran, M., Alfarhood, S., Saha, A.K., Che, D.: Privacy-preserving on-screen activity tracking and classification in e-learning using federated learning. IEEE Access (2023)
13. PASTA, P.T.: Supporting instructional decision making: Potential of an automatically scored three-dimensional assessment system. https:/ai4stem.org/pasta/ (2023)
14. Porras, J.M., Lara, J.A., Romero, C., Ventura, S.: A case-study comparison of machine learning approaches for predicting student's dropout from multiple online educational entities. Algorithms **16**(12), 554 (2023)
15. Rousi, R., Alanen, H.K., Wilson, A.S.: Data privacy, ethics and education in the era of AI–a university student perspective. In: Proceedings of the Conference on Technology Ethics 2024 (Tethics 2024). RWTH Aachen (2024)
16. Truex, S., et al.: A hybrid approach to privacy-preserving federated learning. In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, pp. 1–11 (2019)
17. Xu, B., Yan, S., Li, S., Du, Y.: A federated transfer learning framework based on heterogeneous domain adaptation for students' grades classification. Appl. Sci. **12**(21), 10711 (2022)

18. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. J. Res. Sci. Teach. **59**(10), 1765–1794 (2022)
19. Zhang, P., Zeng, G., Wang, T., Lu, W.: TinyLlama: An open-source small language model. arXiv preprint arXiv:2401.02385 (2024)
20. Zhang, T., et al.: Enhancing dropout prediction in distributed educational data using learning pattern awareness: a federated learning approach. Mathematics **11**(24), 4977 (2023)
21. Zheng, X., Cai, Z.: Privacy-preserved data sharing towards multiple parties in industrial IoTs. IEEE J. Sel. Areas Commun. **38**(5), 968–979 (2020)

Dude, Where's My Utterance? Evaluating the Effects of Automatic Segmentation and Transcription on CPS Detection

Videep Venkatesha(✉), Mariah Bradford, and Nathaniel Blanchard

Colorado State University, Fort Collins, CO 80523, USA
videep@rams.colostate.edu

Abstract. Collaborative Problem-Solving (CPS) markers capture key aspects of effective teamwork, such as staying on task, avoiding interruptions, and generating constructive ideas. An AI system that reliably detects these markers could help teachers identify when a group is struggling or demonstrating productive collaboration. Such a system requires an automated pipeline composed of multiple components. In this work, we evaluate how CPS detection is impacted by automating two critical components: transcription and speech segmentation. On the public Weights Task Dataset (WTD), we find CPS detection performance with automated transcription and segmentation methods is comparable to human-segmented and manually transcribed data; however, we find the automated segmentation methods reduces the number of utterances by 26.5%, impacting the granularity of the data. We discuss the implications for developing AI-driven tools that support collaborative learning in classrooms.

Keywords: Collaborative learning · Machine Learning · Automated Pipelines

1 Introduction

A widely used approach for assessing group work is through identifying Collaborative Problem-Solving (CPS) markers. CPS is a key 21st-century skill emphasized in frameworks like PISA 2015 [19,20,22]. CPS markers categorize group interactions based on behavioral indicators. These markers, as outlined in [23], are divided into three primary facets: (1) Constructing Shared Knowledge, where students build common ground and exchange information; (2) Negotiation and Coordination, where they propose solutions, refine ideas, and resolve conflicts; and (3) Maintaining Team Function, which involves regulating interactions and ensuring productive collaboration.

Previous studies have explored CPS in diverse contexts, including work-based learning [16,20,22]. Additionally, [13] introduced computational linguistic methods for automatically characterizing students' CPS skills, demonstrating that

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 144–151, 2025.
https://doi.org/10.1007/978-3-031-99267-4_18

automated approaches could serve as viable alternatives to human-driven methods. [4] further established that CPS markers could be detected using multimodal features in naturalistic, unconstrained group settings, providing foundational work for automated classroom monitoring. However, while their work demonstrated the viability of automated CPS detection broadly, it does not explore the impact and nuances of using a fully automated system.

To achieve fully automated CPS detection, multiple interconnected components must be carefully considered, each contributing significantly to the reliability of the final system. Crucially, these components include segmentation methods and transcription quality, both of which substantially influence downstream classification performance. Figure 1 illustrates this issue: over the same time window, the Oracle segmentation captures two separate utterances, each with distinct content and collaborative function, while the automated system merges them into one. This leads to both label overlap and transcript compression, which can obscure the speaker turns and degrade downstream CPS classification

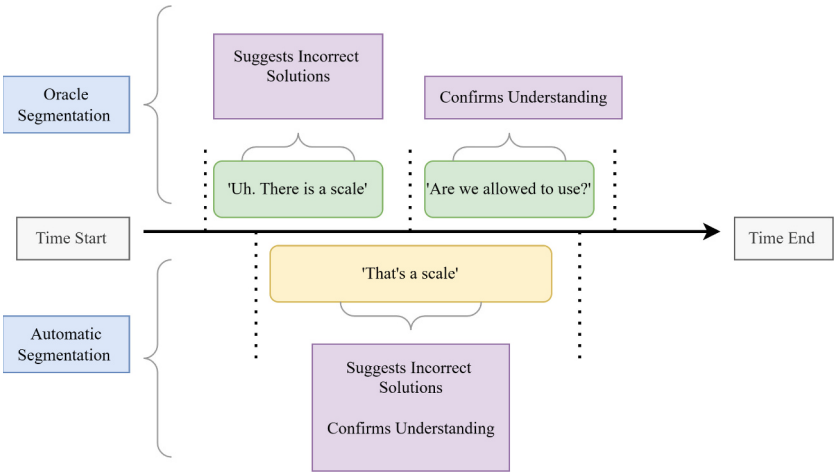


Fig. 1. Comparison of Oracle segmentation (green) and Google automatic segmentation (yellow) over the same time span. Oracle preserves distinct utterances, transcripts, and labels; Google merges them, reducing granularity. (Color figure online)

Segmentation, in this context, refers to the division of continuous group audio into discrete speaker-aligned utterances. The specific impact of using automated segmentation and transcription on the accuracy and granularity of CPS detection remains under-explored in current literature. While Automatic speech recognition (ASR) systems have been extensively evaluated in various contexts [9,15,21,26–28], few studies specifically assess ASR within classroom discourse [2,3]. Cao et al. [6] found that ASR errors severely impacted lexically sensitive tasks but had less effect on discourse-level classification. However, their study

focused broadly on discourse modeling rather than explicitly on CPS, which requires a more nuanced, finer-grained analysis of collaborative interactions.

In this paper, we directly address the research gap by investigating the following research question: How does the use of automated transcription and segmentation affect the detection of CPS markers in small group dialogues? Our findings indicate that fully automated CPS detection pipelines, utilizing ASR-based transcription and segmentation, do not significantly degrade performance compared to human-segmented and manually transcribed data. However, this comes with the caveat that automated methods may not capture CPS behaviors at a granular level, potentially merging distinct utterances and reducing interpretability.

2 Methodology

2.1 Dataset: The Weights Task Dataset

We utilize the Weights Task Dataset (WTD) [18], which consists of ten triads engaged in a structured problem-solving activity, where participants must determine the weights of a set of blocks using a balance scale. Each interaction is annotated for CPS indicators, following the framework outlined by [23]. This framework includes 19 distinct markers of collaborative behavior, with each of them being mapped to constructing shared knowledge, negotiation/coordination, and maintaining team function. The dataset includes Gold-standard (Oracle) segmentation and transcriptions which were manually annotated to serve as the most accurate reference and Google Automatic Speech Recognition (ASR) transcriptions, which was automatically segmented using Google's Voice Activity Detector.

Feature Extraction. We extracted both linguistic and acoustic modalities from the dataset. We extracted linguistic features using BERT-base-small and prosodic features using openSMILE. These were concatenated to form a compact multimodal representation of each utterance, following prior work [7,10,11].

2.2 Modeling Approach

We employ a supervised learning framework to classify Collaborative Problem Solving (CPS) facets. The classification task is framed as a multi-label classification problem, where each utterance can belong to one or more CPS facets: Constructing Shared Knowledge (Const), Negotiation and Coordination (Neg), and Maintaining Team Function (Maintain).

CPS Indicator Labels. Each utterance was assigned to a CPS facet if it contained any subcategory corresponding to that facet. For example, if an utterance contained both *Suggests appropriate ideas* and *Confirms understanding*, it would be labeled as Const = 1, while Neg = 0 and Maintain = 0, since these two facets

fall under *Constructing Shared Knowledge*. This ensured that all relevant utterances contributed to their respective high-level categories. The Google segments inherited the labels of all of the oracle segments that fell under the time span of the Google segment.

Classification Models. We experimented with Random-Forest [5], and AdaBoost [14] following the methodology in [4] to evaluate the effectiveness of different classification models. Deep learning models were not considered, as the size of the dataset was insufficient to support their effective training without overfitting. Hyperparameter tuning was performed using *Hyperopt*, optimizing for *average AUROC scores* over 500 iterations [1]. The search was conducted using the Oracle-Segmented and Oracle-Transcripts condition, and the best-performing model was then evaluated across all other conditions. The best model identified was a Random Forest classifier with criterion set to `entropy`, max features set to `None`, and number of estimators set to `148`. We employ leave-one-group-out cross-validation to evaluate model performance. This is a common technique used to ensure robustness of the machine learning model [8]. We assess model performance using AUROC (Area Under the Receiver Operating Characteristic Curve). AUROC values range from 0 to 1, where 0.5 is random guessing and 1 is perfect [12]. Aditionally, we also calulate the average Precision which measures the proportion of predicted positive labels that are actually correct, and Recall which measures the proportion of actual positive labels that were correctly identified, of the 3 CPS facets, for all conditions.

3 Results

Table 1 presents AUROC scores for different combinations of segmentation and transcription methods. As expected, the highest AUROC is achieved with Oracle transcripts and segmentation (average AUROC = 0.744), while the fully automated condition (Google segmentation and transcription) yields the lowest performance (average AUROC = 0.679). Notably, using Google segmentatizon with Oracle transcripts nearly matches the OracleâĂŞOracle performance (0.740 vs. 0.744). Exploring each condition further provides deeper insight into the nature of these effects. Table 2 presents the average precision and recall across all CPS facets.

We observe that automated transcription substantially degrades precision (from 0.704 to 0.528) and recall (from 0.297 to 0.246) when manual high-quality segmentation is used. This suggests that transcription quality plays a critical role in preserving classifier performance. Automated segmentation has a more nuanced impact: when using oracle transcripts, precision drops slightly (from 0.704 to 0.658), but recall improves (from 0.297 to 0.339), suggesting that merged utterances may help capture broader CPS behaviors at the cost of granularity.

Interestingly, the fully automated condition (Google segmentation and transcription) achieves the highest recall (0.342) despite lower precision (0.601), indicating potential utility for real-time applications that prioritize broader

Table 1. AUROC scores by segmentation and transcription method. Highest values in bold.

Seg.	Tran.	Const(±SD)	Neg(±SD)	Maintain(±PSD)	Average
Oracle	Oracle	**0.758 (0.044)**	**0.765 (0.045)**	**0.720 (0.045)**	**0.744**
Oracle	Google	0.721 (0.029)	0.720 (0.049)	0.653 (0.031)	0.698
Google	Oracle	0.754 (0.066)	0.757 (0.047)	0.699 (0.061)	0.740
Google	Google	0.718 (0.079)	0.705 (0.071)	0.616 (0.078)	0.679

Table 2. Average precision and recall across CPS facets. Highest values in bold.

Seg.	Tran.	Avg. Precision	Avg. Recall
Oracle	Oracle	**0.704**	0.297
Oracle	Google	0.528	0.246
Google	Oracle	0.658	0.339
Google	Google	0.601	**0.342**

detection. The poorest-performing condition is Oracle segmentation with Google transcription—likely due to a mismatch between clean segment boundaries and noisy transcriptions.

4 Discussion

A striking finding of this study is that the use of automated segmentation and ASR-based transcription does not seem to significantly degrade CPS classification performance compared to oracle transcriptions and human-segmented data. This result is particularly promising, as it suggests that fully automated pipelines can still achieve a reasonable level of CPS detection accuracy. However, while CPS classification performance remains high with automatic segmentation, we found the number of instances decreased, as shown in Table 3. This is due to multiple oracle-segmented utterances being merged into a single segment under Google's system. We observed 518 such instances where multiple oracle segments were combined, with 18 cases merging more than two oracle segments. The underlying distribution of labels reveals important limitations of automated segmentation.

Table 3. Number of utterances and CPS labels under Oracle vs. Google segmentation.

	Utterances	Const	Neg	Maintain
Oracle Segments	2482	906	866	391
Google Segments	1824	664	642	328

Merging distinct utterances into a single segment leads to multi-label segments and loss of granularity as seen in Fig. 1, where the distinction between the two CPS facets is blurred. The merging of segments may obscure important micro-level interactions, potentially limiting an AI-driven system's ability to make fine-grained distinctions in collaborative behaviors, and other downstream tasks [24,25]. While automated segmentation allows for efficient large-scale CPS classification, it sacrifices interpretability and label precision.

Beyond technical feasibility, the ability to automatically detect CPS behaviors has meaningful implications for classroom practice. If integrated into a teacher-facing dashboard, such a system could offer real-time or post-hoc summaries of group interactions, helping educators assess students' collaboration skills. However, for such insights to be actionable, teachers need to know not just what behavior occurred, but who said what and when. Segmentation choices directly affect this interpretability: over-segmentation can fragment meaningful behaviors, while under-segmentation can blur speaker turns and behavior distinctions. Understanding this trade-off is crucial for real-world deployment. While our study focuses on Google's segmentation system, which tends to under-segment, it does not examine other ASR pipelines that might over-segment or exhibit different boundary behavior. Exploring a wider range of segmentation strategies could yield substantially different outcomes. A single dataset (WTD) [17] was chosen for its annotated CPS labels and controlled task structure. While this ensures consistency, it limits generalizability to other collaborative contexts (e.g., open-ended tasks or diverse age groups). Additionally, our evaluation maps oracle CPS labels onto automatically segmented transcripts. This decision simplifies comparison across segmentation methods but different mapping strategies could result in varying classification performance.

Conclusion. Our study demonstrates that fully automated pipelines for CPS detection utilizing ASR-based transcription and segmentation can achieve promising levels of performance with only minimal degradation compared to oracle transcriptions. This indicates that automated CPS monitoring systems in classroom settings are feasible. However, our study also highlights the need for more precise segmentation methods or post-processing strategies are needed to preserve fine-grained interaction. This work highlights the potential and limitations of AI-driven approaches for modeling CPS.

Acknowledgements. 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. Bergstra, J., Yamins, D., Cox, D.: Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th International Conference on Machine Learning (ICML), pp. 115–123 (2013)
2. Blanchard, N., et al.: A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 23–33. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19773-9_3
3. Bradford, M., Hansen, P., Ross, J.B., Krishnaswamy, N., Blanchard, N.: A deep dive into microphone hardware for recording collaborative group work. In: Educational Data Mining Conference. Zenodo (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. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
6. Cao, J., et al.: A comparative analysis of automatic speech recognition errors in small group classroom discourse. In: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), pp. 250–261. ACM, Limassol, Cyprus (2023). https://doi.org/10.1145/3565472.3595606
7. Devlin, J.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Springer (1993)
9. Evanini, K., Higgins, D., Zechner, K.: Automated speech scoring for non-native middle school students with multiple task types. In: Proceedings of Interspeech (2013)
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. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010)
12. Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
13. Flor, M., Yoon, S.Y., Liu, O.L., Wagner, M.: Automated classification of collaborative problem solving interactions in simulated science tasks. ETS Res. Rep. Series **2016**(1), 1–12 (2016)
14. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997). https://doi.org/10.1006/jcss.1997.1504
15. Gales, M.J., Young, S.J.: The application of hidden markov models in speech recognition. Found. Trends Sig. Process. **1**(3), 195–304 (2008)
16. Jensen, A.K., et al.: Collaborative problem solving: a pedagogy for workplace relevance. Nord. J. Vocat. Educ. Training **13**(2), 45–73 (2023)
17. Khebour, I., et al.: The Weights Task Dataset: A Multimodal Dataset of Collaboration in a Situated Task (2023)
18. Khebour, I., et al.: When text and speech are not enough: a multimodal dataset of collaboration in a situated task. J. open Humanit. Data **10**(1) (2024)

19. Kids, T.: Collaborative problem solving. https://thinkkids.org/Schools/ (2025)
20. Klieme, E., Hartig, J., Rauch, D., Blum, W.: Assessing collaborative problem solving: an overview of the PISA 2015 assessment framework. In: Collaborative problem solving: An educational perspective, pp. 31–53. Springer (2016)
21. Manuvinakurike, R., DeVault, D.: Using ASR word confusion networks for modeling decisions in spoken dialogue systems. In: Proceedings of Interspeech (2015)
22. OECD: Pisa 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy. Paris: OECD Publishing (2017)
23. Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N., D'Mello, S.: Towards a generalized competency model of collaborative problem solving. Comput. Educ. **143**, 103672 (2020)
24. Venkatesha, V., et al.: Propositional extraction from natural speech in small group collaborative tasks. In: PaaÃÅÿen, B., Epp, C.D. (eds.) Proceedings of the 17th International Conference on Educational Data Mining, pp. 169–180. International Educational Data Mining Society, Atlanta, Georgia, USA (2024). https://doi.org/10.5281/zenodo.12729792
25. Venkatesha, V., et al.: Propositional extraction from collaborative naturalistic dialogues. J. Educ. Data Mining **17**(1), 183–216 (2025)
26. Yoon, S.Y., Xue, Y., Warschauer, M.: Speech-to-text for literacy: how ASR errors affect educational applications. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5059–5070 (2020)
27. Yu, D., Deng, L.: Automatic Speech Recognition: A Deep Learning Approach. Springer (2014)
28. Zechner, K., Evanini, K., Yoon, S.Y., Wang, X.: The challenges of ASR in automated speaking assessment. In: Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE) (2015)



VTutor: An Animated Pedagogical Agent SDK that Provide Real Time Multi-model Feedback

Eason Chen[1(✉)], Chenyu Lin[2], Yu-Kai Huang[1], Xinyi Tang[1], Aprille Xi[1], Jionghao Lin[1,3], and Kenneth Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh, USA
easonc13@cmu.edu
[2] New York University, New York, USA
[3] The University of Hong Kong, Pok Fu Lam, Hong Kong

Abstract. Pedagogical Agents (PAs) show significant potential for boosting student engagement and learning outcomes by providing adaptive, on-demand support in educational contexts. However, existing PA solutions are often hampered by pre-scripted dialogue, unnatural animations, uncanny visual realism, and high development costs. To address these gaps, we introduce VTutor, an open-source SDK leveraging lightweight WebGL, Unity, and JavaScript frameworks. VTutor receives text outputs from a large language model (LLM), converts them into audio via text-to-speech, and then renders a real-time, lip-synced pedagogical agent (PA) for immediate, large-scale deployment on web-based learning platforms. By providing on-demand, personalized feedback, VTutor strengthens students' motivation and deepens their engagement with instructional material. Using an anime-like aesthetic, VTutor alleviates the uncanny valley effect, allowing learners to engage with expressive yet comfortably stylized characters. Our evaluation with 50 participants revealed that VTutor significantly outperforms the existing talking-head approaches (e.g., SadTalker) on perceived synchronization accuracy, naturalness, emotional expressiveness, and overall preference. As an open-source project, VTutor welcomes community-driven contributions—from novel character designs to specialized showcases of pedagogical agent applications—that fuel ongoing innovation in AI-enhanced education. By providing an accessible, customizable, and learner-centered PA solution, VTutor aims to elevate human-AI interaction experience in education fields, ultimately broadening the impact of AI in learning contexts. The demo link to VTutor is at https://vtutor-aied25.vercel.app.

Keywords: Animated Pedagogical Agents · Generative AI · Talking Head

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 152–159, 2025.
https://doi.org/10.1007/978-3-031-99267-4_19

1 Introduction and Background

AI in Education (AIED) aims to enhance learning processes and outcomes by leveraging artificial intelligence techniques [22] such as intelligent tutoring systems [24], adaptive assessments [20], programming [3], and chatbot [2]. While large language models (LLMs) now excel at generating personalized content and responding to learners' open-ended queries [1,4,6,16,26], many of these systems are constrained by purely text-based or voice-based interfaces. *Pedagogical Agents* (PAs) offers an opportunity for a more human-like interaction experience with LLMs [15,17,23], delivering immediate, adaptive guidance via facial expressions, gestures, and voice [10,11,18,25].

However, existing PA frameworks often rely on scripted interactions [9,21], utilize static or non-synchronized facial expressions [15], high development cost, and long latency [17,27]. These constraints lead to interaction delays, unnatural facial movements, and overall lower engagement, thereby limiting the development of dynamic, learner-centered AI tutoring experiences. As a result, these PAs cannot reliably provide responsive feedback, emotional support, and adaptive scaffolding—key features that enhance the tutoring process. Moreover, when a PA design becomes overly humanlike, it risks triggering the "uncanny valley" phenomenon [8,23], wherein users feel discomfort toward entities that appear almost, but not exactly, human. This effect can discourage learners from interacting with the PA, undercutting its potential to facilitate meaningful learning outcomes. Consequently, there is a growing need for an integrated solution that aligns robust LLM-driven dialogue capabilities with visually appealing, synchronized, and scalable PA tools, all while avoiding the uncanny valley effect to preserve learner engagement and trust.

Meanwhile, virtual avatars have increasingly become a presence in real-world applications, particularly in the entertainment field, where many content creators opt to engage with their audiences using digital personas instead of revealing their real faces. These creators, known as *VTubers* [7,13], leverage face tracking and motion capture technologies to animate their avatars in real-time by 3D engine like Unity, enabling expressive and interactive performances. The aesthetic of VTuber avatars is heavily influenced by anime culture, incorporating cuteness and charm elements that make them visually appealing and emotionally engaging for viewers [7,13]. These stylistic choices not only enhance relatability but also circumvent the uncanny valley effect, maintaining user comfort and engagement.

Our Contribution: This paper introduces **VTutor**, an open-source Software Development Kit (SDK) designed to streamline the development and deployment of animated pedagogical agents (APAs) in educational settings. VTutor enhances human-AI interaction by enabling real-time, engaging, and emotionally expressive AI-driven tutors. Specifically, VTutor contributes:

1. **A Generative AI-Driven Interactive Agent:** By integrating output from LLMs with animated pedagogical agents, VTutor facilitates real-time, personalized, and adaptive interaction experiences.
2. **Customizable Characters:** VTutor adopts anime-style characters that strike a balance between realism and abstraction, avoiding the uncanny valley while enhancing engagement across diverse learners.
3. **A Seamless, Scalable, and Developer-Friendly SDK:** The lightweight browser-compatible architecture ensures that VTutor can be effortlessly embedded into existing educational platforms with minimal setup.

By addressing key challenges in traditional APAs—such as lack of real-time adaptability, poor lip synchronization, and high computational overhead—VTutor paves the way for more immersive and accessible AI-driven educational tools (Fig. 1).

2 VTutor System Implementation

VTutor is composed of the following components:

(1) Large Language Model Integration. VTutor presents AI-generated text responses by integrating various LLMs (e.g., GPT, Claude, Deepseek). Educators can customize prompts or knowledge bases to align PA interactions with learning objectives, enabling flexible, contextually adaptive dialogues.

(2) Text-to-Speech (TTS). VTutor accepts audio in `.wav` format from major TTS services (e.g., Azure, ElevenLabs). This enables high-quality, natural speech that resonates with learners. During run-time, the system retrieves newly generated text from the LLM, converts it to audio via a TTS API, and sends it to the Unity engine for real-time facial expression and mouth rendering.

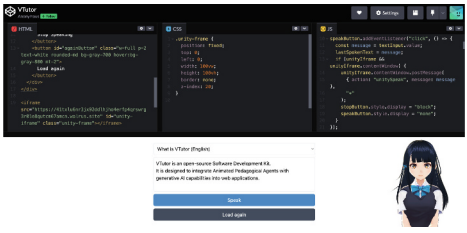


Fig. 1. Screenshot of the VTutor SDK demo at https://vtutor-aied25.vercel.app/sdk, the few lines of codes in the image are what developers need to do to add VTutor to their own website. VTutor is the animated avatar presented at the bottom right.

(3) LipSync Module. A Unity lip-sync library [14] analyzes incoming audio to detect phoneme sequences. Corresponding mouth-blend shapes are applied to a 3D or 2D avatar, achieving near real-time alignment with an average latency of under one second for short utterances. This approach yields more accurate and smooth mouth movements compared to naive keyframe or jaw-only animations.

(4) WebGL-Based Rendering. VTutor packages the Unity environment into a WebGL build, allowing easy embedding in any web page by iframe. Communication between the web page and the embedded PA occurs via an open-sourced software development kit (SDK) library, enabling dynamic control and fine-grained interactivity (e.g., changing gestures, expressions, or switching avatars).

3 Evaluation Method and Results

We invited participants to compare VTutor with talking-head pedagogical agents generated using SadTalker [27], following the methodology described by [17]. To minimize potential appearance biases, we used the same VTutor agent design to create the SadTalker talking head. The full evaluation materials can be found at https://vtutor-aied25.vercel.app/evaluation.

Our experiment received approval from the Institutional Review Board (IRB). We recruited 50 participants for evaluation through convenience sampling via Prolific. The participants were between 20 and 50 years old, consisting of 30 males and 20 females. We conducted the experiment online, where participants watched the videos on YouTube and responded to the questions using Google Forms.

The videos featured VTutor and Sad Talker speaking, but throughout the entire experiment, they were referred to only as Agent A and Agent B. Among them, 25 random participants viewed VTutor on the left as Agent A and SadTalker on the right as Agent B, while the remaining 25 participants saw the opposite arrangement. After watching the videos, participants were asked which agent they preferred and were asked to rate the following dimensions for each agent on a scale from 1 to 7, and answered open-ended questions about their opinions.

- **General Preference Score**: Preference rating between the two agents.
- **Sync Accuracy**: Evaluates the alignment between facial expressions, lip movements, and audio.
- **Naturalness**: Assesses whether the video appears natural and not rigid.
- **Emotional Expression**: Determines whether the facial expressions effectively convey the agent's emotions.
- **Visual Coherence**: Examines the coordination between lip movements and other facial features.

For quantitative data analysis, we used the Chi-square test for user preference. For the user rating, we conducted an independent samples t-test. We posed

Table 1. Comparison of VTutor and SadTalker on different evaluation metrics

Question	VTutor Mean (SD)	SadTalker Mean (SD)	t statistics	p-value	Cohen's d
General Preference Score	5.00 (1.75)	3.62 (1.51)	4.2218	0.0001	0.8444
Sync Accuracy	4.58 (1.82)	3.66 (1.49)	2.7642	0.0068	0.5528
Naturalness	4.12 (2.04)	3.10 (1.67)	2.7390	0.0073	0.5478
Emotional Expression	4.38 (1.94)	3.30 (1.58)	3.0546	0.0029	0.6109
Visual Coherence	4.56 (1.64)	3.64 (1.63)	2.8142	0.0059	0.5628

open-ended questions about the two agents to gather qualitative insights through thematic analysis [12], which is well-suited for exploratory studies without the need of agreement scores as suggested by [19], as our main goal was to capture a broad range of perspectives rather than measure inter-rater reliability.

3.1 Quantitative User Preference Evaluation Results

The participants exhibited a strong preference for the VTutor version. Among the 50 participants, 36 preferred VTutor, while only 14 favored Sad Talker. The statistical analysis yielded a chi-square statistic of $\chi^2(1, N = 50) = 9.68$, with a p-value of $p = .0019$ and Cramer's V of $V = .44$, indicating a moderate to strong effect size. Moreover, as shown in Table 1, VTutor demonstrated a significant advantage across all evaluation dimensions ($p < .01$). These results suggest that VTutor, utilizing the Unity technology framework, provides a more engaging, synchronized, natural, emotionally expressive, and visually coherent PA compared to the Sad Talker-generated talking head.

3.2 Qualitative Insights

Concerns About SadTalker's Uncanny Valley. Two participants explicitly mentioned being disturbed by SadTalker's appearance. One stated, *"[SadTalker] was a bit unsettling and looked strange"* (P35), while another directly commented, *"[SadTalker] look weird and uncanny"* (P36). In total, many participants (e.g., P24, P28, P35, P41, P45) described SadTalker as "unsettling," "creepy," or "weird," suggesting an uncomfortable, uncanny-valley-like experience.

Positive Feedback and Remaining Issues on VTutor. By contrast, VTutor received favorable remarks from several participants, particularly for its clarity and color. One noted, *"[VTutor] was overall brighter and easier to see. Their lip movements also did a better job of matching the audio."* (P50). Meanwhile, another participant appreciated how its stylized mouth avoided some uncanny pitfalls: *"It didn't look 'natural,' but [VTutor] was a lot closer to an anime character in the mouth. That in itself made it feel less creepy.'* (P41). Despite these positives, a few participants still felt VTutor's facial expression animations were also "weird" or "unnatural" (P28, P40, P45). While its anime-like appearance seemed to alleviate the uncanny valley, further refinements on VTutor avatar are needed to achieve a more convincing, natural-looking facial expression.

3.3 Computational Time and Latency

In our experiment, SadTalker required 143 s to generate a 7-second video on an NVIDIA Titan XP GPU, whereas VTutor produced the animation within one second via a browser environment. This efficiency is critical for real-time interactive pedagogical agents. Moreover, VTutor can be integrated with just an HTML iframe and requires no specialized backend to compute the animation rendering, underscoring its lightweight and scalability in deploying APA at scale.

4 Discussion

Advancing Animated Pedagogical Agents. VTutor redefines the design and capabilities of animated Pedagogical Agents (PAs) by leveraging generative AI and real-time animation. Existing solutions often suffer from either scripted dialogues or rigid, non-synchronized facial expressions. In contrast, VTutor's modular pipeline integrates large language models (LLMs), text-to-speech (TTS), and real-time lip synchronization to provide personalized, adaptive feedback in a more fluid and engaging manner. This dynamic synergy ensures that the agent's speech, mouth movements, and expressions are meticulously aligned, mitigating the robotic feel or delay that commonly characterizes earlier systems.

Beyond technical innovation, VTutor prioritizes the user experience. Our evaluation demonstrates that participants perceive the system to be more natural, well-synchronized, and visually appealing, which will foster deeper engagement. By delivering human-like, contextually responsive guidance, VTutor bolsters the motivational and emotional underpinnings crucial for effective learning.

Anime-Style Avatars and the Uncanny Valley. A key insight emerging from our user study is that anime-like avatars may reduce the uncanny valley effect often encountered with realistic character models. Participants frequently described SadTalker's talking-head approaches as "unsettling" or "creepy," highlighting how slight deviations from human realism can evoke discomfort. In contrast, VTutor's animated style characters, with simpler facial features and intentional abstraction, were praised for feeling more natural and less uncanny.

This suggests that designing PAs with an animated style is not merely an artistic choice but a strategic one to maintain user engagement. By balancing expressive detail with clear abstraction, anime-like avatars avoid the pitfalls of near-realistic facial features that sometimes trigger negative emotional responses. This finding has broad implications for PA design, emphasizing the value of recognizable yet stylized models that preserve learners' comfort and trust.

Limitations and Future Directions. Despite VTutor's promising results, VTutor's animations still have room to improve. Incorporating refined 3D models or advanced animations could help produce more lifelike interactions. Additionally, VTutor's current Unity WebGL build of approximately 115 MB can result in long initial loading times. Future efforts will focus on package optimization and progressive loading to enhance performance and scalability.

5 Conclusion and Potential Applications

VTutor seamlessly combines large language models (LLMs), text-to-speech, and lip-sync to render adaptive, anime-style pedagogical agents in a lightweight browser environment. Evaluation results show that VTutor outperforms conventional talking-head approaches in lip synchronization, naturalness, and emotional expressiveness. By offering an open-source, customizable toolkit, VTutor paves the way for more engaging AI-driven pedagogical agents interaction experiences.

Leveraging its user-friendly, real-time, and scalable architecture, VTutor could provide immediate and adaptive feedback by pedagogical agents using the output from LLMs and thus effectively addresses a wide range of educational needs. From high-impact STEM tutoring and language learning peers [5] to personalized workforce skill development and learner social support, VTutor has the potential to enrich human-AI interactions across various educational domains.

References

1. Ali, O., Murray, P.A., Momin, M., Dwivedi, Y.K., Malik, T.: The effects of artificial intelligence applications in educational settings: challenges and strategies. Technol. Forecast. Soc. Chang. **199**, 123076 (2024)
2. Chen, E.: The effect of multiple replies for natural language generation chatbots. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1–5 (2022)
3. Chen, E., Huang, R., Chen, H.S., Tseng, Y.H., Li, L.Y.: Gptutor: a ChatGPT-powered programming tool for code explanation. In: International Conference on Artificial Intelligence in Education, pp. 321–327. Springer (2023)
4. Chen, E., Lee, J.E., Lin, J., Koedinger, K.: Gptutor: great personalized tutor with large language models for personalized learning content generation. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale, pp. 539–541 (2024)
5. Chen, E., et al.: Vtutor for high-impact tutoring at scale: managing engagement and real-time multi-screen monitoring with p2p connections. arXiv preprint arXiv:2505.07736 (2025)
6. Chen, E., Wang, D., Xu, L., Cao, C., Fang, X., Lin, J.: A systematic review on prompt engineering in large language models for k-12 stem education. arXiv preprint arXiv:2410.11123 (2024)
7. Chinchilla, P., Kim, J.: Vtuber for streamers: exploring the role of social presence in the visual representation of streamers. Commun. Stud. 1–17 (2024)
8. Ciechanowski, L., Przegalinska, A., Magnuski, M., Gloor, P.: In the shades of the uncanny valley: an experimental study of human-chatbot interaction. Futur. Gener. Comput. Syst. **92**, 539–548 (2019)
9. Cramify: Cramify: the new way to cram — AI study guides & More — cramify.ai (2025). https://cramify.ai/. Accessed 23 Jan 2025
10. Davis, R.O., Park, T., Vincent, J.: A systematic narrative review of agent persona on learning outcomes and design variables to enhance personification. J. Res. Technol. Educ. **53**(1), 89–106 (2021). https://doi.org/10.1080/15391523.2020.1830894
11. Domagk, S.: Do pedagogical agents facilitate learner motivation and learning outcomes?: the role of the appeal of agent's appearance and voice. J. Media Psychol. **22**(2), 84–97 (2010). https://doi.org/10.1027/1864-1105/a000011

12. Fereday, J., Muir-Cochrane, E.: Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int. J. Qual. Methods **5**(1), 80–92 (2006)
13. Ferreira, J.C.V., Regis, R.D.D., Gonçalves, P., Diniz, G.R., Tavares, V.P.D.S.C.: Vtuber concept review: the new frontier of virtual entertainment. In: Proceedings of the 24th Symposium on Virtual and Augmented Reality, pp. 83–96 (2022)
14. hecomi: GitHub - hecomi/uLipSync: MFCC-based LipSync plug-in for unity using job system and burst compiler — github.com (2021). https://github.com/hecomi/uLipSync. Accessed 23 Jan 2025
15. Heeyo: Heeyo — heeyo.life (2024). https://heeyo.life/. Accessed 23 Jan 2025
16. Kasneci, E., et al.: ChatGPT for good? on opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
17. Lin, J., Chen, E., Gurung, A., Koedinger, K.R.: Mufin: a framework for automating multimodal feedback generation using generative artificial intelligence. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale, pp. 550–552 (2024)
18. Makransky, G., Wismer, P., Mayer, R.E.: A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. J. Comput. Assist. Learn. (2018)
19. 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**(CSCW), 1–23 (2019)
20. Minn, S.: Ai-assisted knowledge assessment techniques for adaptive learning environments. Comput. Educ. Artif. Intell. **3**, 100050 (2022)
21. Okado, Y., Nye, B.D., Aguirre, A., Swartout, W.: Can virtual agents scale up mentoring?: insights from college students' experiences using the careerfair. ai platform at an American hispanic-serving institution. In: International Conference on Artificial Intelligence in Education, pp. 189–201. Springer (2023)
22. Self, J.: The birth of ijaied. Int. J. Artif. Intell. Educ. **26**(1), 4–12 (2016)
23. Sikström, P., Valentini, C., Sivunen, A., Kärkkäinen, T.: Pedagogical agents communicating and scaffolding students' learning: high school teachers' and students' perspectives. Comput. Educ. **222**, 105140 (2024)
24. Stamper, J., Xiao, R., Hou, X.: Enhancing LLM-based feedback: insights from intelligent tutoring systems and the learning sciences. In: International Conference on Artificial Intelligence in Education, pp. 32–43. Springer (2024)
25. Veletsianos, G.: How do learners respond to pedagogical agents that deliver social-oriented non-task messages? impact on student learning, perceptions, and experiences. Comput. Hum. Behav. **28**(1), 275–283 (2012). https://doi.org/10.1016/j.chb.2011.09.010
26. Wang, S., et al.: Large language models for education: a survey and outlook. arXiv preprint arXiv:2403.18105 (2024)
27. Zhang, W., et al.: Sadtalker: learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8652–8661 (2023)

"Small Device, Big Decision:" Comparing Lightweight LLMs' Computational Performance and Output Quality for AIED Unplugged

Mateus Monteiro[1(✉)], Aristoteles Barros[1], Luiz Rodrigues[1,3], Diego Dermeval[1], Ig Ibert Bittencourt[1], Seiji Isotani[2], and Valmir Macario[4]

[1] Center of Excellence in Social Technologies (NEES) - Federal University of Alagoas, Maceió, Brazil
{mateus.monteiro,aristoteles.barros,luiz.rodrigues,diego.matos, ig.ibert}@nees.ufal.br
[2] Institute of Mathematical and Computer Sciences - University of São Paulo, São Paulo, Brazil
sistotani@icmc.usp.br
[3] Technological Federal University of Paraná (UTFPR) - Apucarana, Curitiba, Brazil
[4] Federal Rural University of Pernambuco (UFRPE) - Pernambuco, Recife, Brazil
valmir.macario@ufrpe.br

Abstract. Deploying Large Language Models (LLMs) in offline educational contexts, particularly on edge devices like smartphones, presents significant computational challenges due to their high resource demands. Addressing this issue, AIED Unplugged proposes using lightweight LLMs optimized for mobile platforms, yet understanding the practical performance-quality trade-offs remains limited. This study investigates these trade-offs through benchmarking six compact LLMs embedded via the Llama.cpp framework into an Android application on a low-resource smartphone. Performance metrics (response time, memory usage, and storage size) were tracked using Sentry, while DeepEval assessed response quality based on relevance, clarity, accuracy, and completeness. The evaluation demonstrated notable trade-offs: lighter models, such as MobileVLM and TinyLlama, exhibited fast inference speeds but lower quality in completeness and accuracy, whereas slightly larger models like Qwen2 and DeepSeek provided enhanced accuracy and completeness but suffered from slower responses and increased memory requirements. These insights underscore the necessity of targeted optimization strategies, including model quantization and hybrid deployments, to effectively balance resource efficiency and educational quality. Ultimately, this research supports equitable access to learning materials for marginalized people by providing insights for AI-powered education in offline situations.

Keywords: Large Language Models · AIED Unplugged · Benchmark · Performance · Quality

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 160–167, 2025.
https://doi.org/10.1007/978-3-031-99267-4_20

1 Introduction

Artificial Intelligence in Education (AIED) offers transformative potential in content generation, feedback, and student engagement [26], yet its benefits often rely on cloud infrastructure—limiting accessibility in low-resource settings where connectivity and advanced hardware are scarce [20]. To mitigate this, offline approaches like AIED Unplugged [14] promote the use of lightweight AI on edge devices. However, the offline feasibility of LLMs remains constrained by their computational demands [30]. Recent work on quantization and distillation has advanced mobile deployment [9,12,24]. This study benchmarks lightweight LLMs on smartphones using Llama.cpp [9], evaluating response time, memory usage, and output quality relative to GPT-4o [13], contributing to digital equity through scalable, cloud-independent educational solutions.

2 Literature Review

The integration of LLMs has expanded the scope of AIED, enabling applications in content generation, tutoring, and assessment [28]. While this marks a major milestone in the field, it also introduces the challenge of ensuring AI supports, rather than replaces, educators' roles. Sustainable use depends on preserving teacher agency and integrating AI in ways that enhance—without overshadowing—human judgment, especially in evaluative contexts. Prior studies have explored how AI can complement educators: Tenorio et al. [25] investigated gamified environments; Dimitri et al. [6] analyzed intelligent tutoring systems; and Jeon and Lee [15] identified four AI roles in classrooms—interlocutor, content provider, assistant, and evaluator—while emphasizing that engagement and ethical responsibility remain with teachers.

Research has also addressed the use of AI in underserved regions, where both infrastructure and pedagogy face constraints [20]. Adaptations include AI tools for Brazilian public schools with limited connectivity [10] and intelligent tutors for low-resource settings [10]. LLMs have enabled advances in language learning, feedback generation, and assessment [19], yet accessibility remains a barrier, especially in the Global South. To address this, Isotani et al. [14] proposed AIED Unplugged—offline AI tools for resource-limited environments. While promising, this raises technical questions about LLM efficiency on low-power devices. Recent efforts show potential: Wang et al. [27] demonstrated multi-task LLMs on mobile, and Santos et al. [18] assessed deployment via the MLC framework [24], noting trade-offs such as increased memory demands and longer response times, depending on model configuration. However, these studies offer limited insight into education-focused benchmarks. This study addresses that gap by positioning LLM evaluation within a low-resource educational context, examining the balance between computational efficiency and response quality under real-world offline conditions.

3 Method

This study investigates the deployment potential of lightweight LLMs within the AIED Unplugged paradigm, which targets offline environments characterized by limited computational resources, such as smartphones. The analysis integrates two dimensions: computational performance, assessed through inference time and resource usage, and response quality, evaluated based on textual coherence under deployment constraints. The implementation leverages Llama.cpp [9], selected for its maturity, active development community, and compatibility with Android environments. This framework was integrated into a custom APK to enable local inference without internet connectivity. While the models evaluated are not explicitly tailored for educational tasks, this benchmarking effort serves as a foundational step toward identifying viable infrastructure for embedding LLMs into mobile educational applications. By understanding the technical boundaries of local inference, this work informs the design of offline intelligent tutoring systems, especially for regions with limited access to connectivity or cloud-based solutions.

3.1 Testbed Environment

To evaluate LLMs in offline educational settings, we used a publicly available dataset [19] aligned with Bloom's taxonomy [3], supporting reproducibility and comparability in AIED research [17]. The dataset contains nearly 800 open-ended questions in high school biology and earth sciences, selected for their curricular relevance and cognitive variety—from topics like global warming to the human skeletal system—capturing typical teacherâĂŞstudent interactions with LLMs. Given the original 8,000-response volume, full evaluation proved unfeasible under resource constraints, leading us to sample 180 biology questions (30 per Bloom level) to maintain representativeness and thematic focus[1].

We selected mobile-optimized models targeting low-cost GPUs and underpowered devices [11], chosen for Android compatibility and efficient memory usage. Their storage and RAM requirements—summarized in the model specifications table based on Hugging Face data [4]—indicate their suitability for deployment on resource-constrained hardware. DeepSeek-R1-Distill-Qwen-1.5B-Q8_0, included for its reasoning capabilities, offers coherent output at modest computational cost [5]. Using Llama.cpp [9], pre-quantized `gguf` models [8] were embedded via Android NDK into a Kotlin app, configured for C/C++ inference with the prompt "*Provide brief and precise answers.*" Monitoring was handled by Sentry [2], with testing conducted on a $239 Samsung Galaxy A34 (Android 14), chosen for its global presence and realistic deployment profile [22]. The LLMs ran reliably despite hardware constraints, supported by Llama.cpp's lightweight architecture.

[1] The dataset's subset used in this study is available from this link.

3.2 Evaluation Analysis

The evaluation considered two complementary dimensions—computational performance and output quality—using descriptive and inferential statistics to assess six models[2]. Computational performance was measured through Response Time, Memory Usage, and Storage Utilization, with real-time monitoring via Sentry [2]. Output quality was assessed using the DeepEval framework [23], which scores Relevance, Clarity, Accuracy, and Completeness on a 0âĂŞ10 scale through techniques such as semantic similarity and readability analysis. An Overall Pass Rate was calculated based on outputs exceeding a quality threshold, using GPT-4o as reference [13]. Descriptive statistics (Mean, SD) summarized metric distributions, while ANOVA and Post Hoc tests identified significant differences across models [21]. Correlation analysis explored interactions between computational and qualitative metrics, revealing trade-offs. The evaluation criteria included: **Relevance**, the semantic and contextual alignment between prompt and output, measured via cosine similarity; **Clarity**, the ease of understanding based on readability and structural flow; **Accuracy**, the correctness of content relative to reference responses using semantic similarity; and **Completeness**, the extent to which key information from the reference is covered, assessed by lexical overlap. This comprehensive approach ensures that model outputs are not only technically adequate but also meet pedagogical expectations [16].

4 Results

As detailed in Table 1, the analysis of computational performance across models reveals an average response time of 41.2 s (SD = 46.5); MobileVLM was the fastest (12.66 s) and DeepSeek the slowest (140.9 s). Memory usage averaged 2631 MB (SD = 231), ranging from 2236 MB (DeepSeek) to 3028 MB (RedPajama), while storage remained constant due to fixed model sizes. Table 2 shows output quality results: Qwen2 had the highest accuracy (8.23) and completeness (3.63), making it ideal for tasks requiring precision. DeepSeek also scored well in accuracy (7.58) and relevance (7.49) but had lower completeness (3.15), while Phi led in clarity (4.69) yet scored lower in completeness (2.89). MobileVLM, despite being the fastest, had the lowest relevance (6.46) and completeness (2.21), consistent with its lightweight design. TinyLlama and RedPajama showed moderate results, with TinyLlama slightly better in relevance (6.85) but lower in clarity (3.47), and RedPajama balanced across relevance (7.00), clarity (3.62), and completeness (2.93). Metric relationships indicated a weak correlation between response time and completeness ($r = 0.150$) and a moderate negative correlation between memory usage and response time ($r = -0.558$). Although the correlation between speed and accuracy was low ($r = 0.129$), a clear pattern emerged—faster models like MobileVLM and TinyLlama scored lower in completeness, whereas slower ones like Qwen2 and DeepSeek performed better, sug-

[2] Selected models' storage size (SS) and estimated RAM usage (RAM) in GB are available from this link.

gesting completeness depends more on architecture than speed. Accuracy correlated strongly with relevance ($r = 0.861$) and completeness ($r = 0.705$), while clarity showed weak negative correlations with both ($r = -0.168$, $r = -0.052$), indicating a possible trade-off between clarity and content depth.

Table 1. Performance Metrics for Evaluated LLMs (Mean)

Model	Response Time (s)	Memory Usage (MB)	Storage Usage (MB)
DeepSeek	140.9 (SD = 26.41)	2310 (SD = 19.69)	1935.36 (SD = N/A)
MobileVLM	12.7 (SD = 2.38)	2735 (SD = 9.88)	792 (SD = N/A)
Phi	22.5 (SD = 12.44)	2940 (SD = 35.7)	1198.08 (SD = N/A)
Qwen2	32.5 (SD = 1.07)	2855 (SD = 45.9)	986 (SD = N/A)
Redpajama	33.3 (SD = 1.10)	2855 (SD = 73.3)	1832.96 (SD = N/A)
Tinyllama	17.2 (SD = 1.17)	2497 (SD = 11.72)	2252.8 (SD = N/A)
ANOVA	$F = 5.51 \times 10^5$; $p < 0.001$	$F = 1.30 \times 10^6$; $p < 0.001$	N/A

DeepSeek: DeepSeek-R1-Distill-Qwen-1.5B-Q8_0; N/A: Not Applicable

Table 2. Evaluation Metrics for LLMs Based on Qualitative Factors (Mean).

Model	Relevance	Clarity	Accuracy	Completeness
DeepSeek	7.49 (SD = 0.65)	3.84 (SD = 0.72)	7.58 (SD = 0.75)	3.15 (SD = 0.68)
MobileVLM	6.46 (SD = 2.17)	3.25 (SD = 1.72)	6.28 (SD = 2.15)	2.21 (SD = 1.19)
Phi	7.72 (SD = 1.30)	4.69 (SD = 2.30)	7.64 (SD = 1.33)	2.89 (SD = 0.99)
Qwen2	7.91 (SD = 0.92)	3.09 (SD = 0.95)	8.23 (SD = 1)	3.63 (SD = 0.89)
Redpajama	7.00 (SD = 1.48)	3.62 (SD = 1.25)	7.19 (SD = 1.44)	2.93 (SD = 0.88)
Tinyllama	6.85 (SD = 2.61)	3.47 (SD = 1.54)	6.75 (SD = 2.41)	2.78 (SD = 1.04)
ANOVA	F = 7521.60; p < 0.001	F = 5298.11; p < 0.001	F = 13940.74; p < 0.001	F = 14281.47; p < 0.001

DeepSeek: DeepSeek-R1-Distill-Qwen-1.5B-Q8_0

The comparative evaluation of output quality reveals clear trade-offs shaped by each model's architecture[3]. DeepSeek achieved high accuracy (M = 7.58), outperforming MobileVLM (M = 6.28, $g = 0.81$) and TinyLlama (M = 5.32, $g = 1.27$), though slightly trailing Qwen2 (M = 8.24, $g = -0.74$). A similar pattern was found for relevance: DeepSeek (M = 7.49) exceeded MobileVLM (M = 6.46, $g = 0.64$) but was outperformed by Qwen2 (M = 7.91, $g = -0.53$), suggesting Qwen2 may be more suitable for tasks requiring higher precision. In computational terms, DeepSeek incurred substantial latency (M = 140.93 s), far above

[3] The detailed results of the post-hoc analyses presented in this study, building upon the significant ANOVA results shown in Tables 1 and 2, as well as the complete correlation matrix, are available from this link.

MobileVLM (M = 12.67 s, $g = 6.86$) and Phi (M = 22.47 s, $g = 5.75$), reflecting a clear trade-off between output quality and response time. While DeepSeek used less memory (M = 2310.59 MB) than MobileVLM (M = 2735.04 MB, $g = -27.33$) and Qwen2 (M = 2497.50 MB, $g = -5.32$), the gain was insufficient to offset its slower performance. A further contrast appeared between clarity and completeness: Phi achieved greater clarity (M = 4.69) than DeepSeek (M = 3.84, $g = -0.50$), whereas DeepSeek delivered more complete answers (M = 3.15 vs. 2.89, $g = 0.30$). These results underscore the need to align model selection with task requirements, balancing linguistic depth against computational constraints.

5 Discussion and Implications

Considering the findings, models with shorter response times—such as MobileVLM and TinyLlama—tended to yield lower completeness and accuracy, underscoring the challenge of balancing speed and output depth. Conversely, Qwen2 and DeepSeek delivered more accurate and complete responses at the expense of longer inference times and higher memory usage. These trade-offs suggest that lightweight models suit real-time applications, while tasks demanding precision may require more resource-intensive options, reinforcing the importance of aligning model choice with specific computational constraints and educational objectives.

The correlation analysis illustrates how performance metrics interact. Clarity correlated weakly and negatively with both accuracy and completeness, while accuracy and relevance showed a strong positive relationship, indicating that more precise models tend to maintain contextual consistency. The weak correlation between response time and completeness suggests that response quality depends more on architecture and parameter tuning than speed alone. These patterns highlight the complexity of LLM optimization, where gains in one area may compromise others. Fine-tuning and adaptive methods—such as quantization, distillation, and selective offloading—can help mitigate trade-offs [7]. Task-specific deployment should guide model selection: AI tutors must balance clarity and completeness to support student comprehension, while assistive tools in classrooms—such as summarization aids or reading companions—may prioritize faster response times and contextual precision [29]. Hybrid strategies like progressive disclosure and dynamic scaling may further adapt responses to learner profiles and task complexity, provided they are supported by broader evaluations in diverse pedagogical scenarios.

This study has limitations. The evaluation was conducted on a single device (Samsung Galaxy A34), limiting generalizability across hardware. To better reflect real-world classroom conditions, future studies should include a broader set of mobile and edge devices, capturing performance variation across different hardware profiles. The dataset size was constrained by resources, reducing the robustness of statistical conclusions; larger datasets could offer more reliable insights into LLM performance. Additionally, only lightweight models were tested, excluding more powerful options like Llama-3.3 [1], which could provide

valuable comparative data. Lastly, metrics such as CPU and battery usage were not included but are relevant for assessing feasibility in sustained classroom use. Addressing these gaps in future work will strengthen understanding of the trade-offs between efficiency and response quality in offline AI-powered education, and support the design of context-aware, device-adaptive educational tools that meet the needs of diverse learners.

References

1. AI, M.: Llama 3.3-70b instruct (2025). https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed 14 Feb 2025
2. Albuquerque, C., Correia, F.F.: Deployment tracking and exception tracking: monitoring design patterns for cloud-native applications. In: EuroPLoP 2023, Proceedings of the 28th European Conference on Pattern Languages of Programs. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3628034.3628038
3. Anderson, L.W., Sosniak, L.A.: Bloom's Taxonomy. University of Chicago Press Chicago, IL, USA (1994)
4. Community, H.F.: Open LLM leaderboard (2024). https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/. Accessed 14 Feb 2025
5. DeepSeek-AI: Deepseek-r1: incentivizing reasoning capability in LLMs via reinforcement learning (2025). https://arxiv.org/abs/2501.12948
6. Di Mitri, D., Schneider, J., Drachsler, H.: Keep me in the loop: real-time feedback with multimodal data. Int. J. Artif. Intell. Educ. **32**(4), 1093–1118 (2022)
7. Dubiel, M., Barghouti, Y., Kudryavtseva, K., Leiva, L.A.: On-device query intent prediction with lightweight LLMs to support ubiquitous conversations. Sci. Rep. **14**(1), 12731 (2024)
8. Face, H.: Hugging face - the AI community building the future (2024). https://huggingface.co. Accessed 14 Feb 2025
9. Gerganov, G.: Llama.cpp (2023). https://github.com/ggerganov/llama.cpp
10. Guerino, G., et al.: We see you: understanding math teachers from brazilian public schools to design equitable educational technology. Revista Brasileira de Informática na Educação (RBIE) **32**, 336–358 (2024). https://doi.org/10.5753/rbie.2024.3649, https://sol.sbc.org.br/journals/index.php/rbie/article/view/3649
11. Hannibal046: Awesome LLM: a curated list of large language model resources (2024). https://github.com/Hannibal046/Awesome-LLM. Accessed 14 Feb 2025
12. Inc., P.: picollm — towards optimal LLM quantization (2024). http://picovoice.ai/blog/picollm-towards-optimal-llm-quantization/
13. Islam, R., Moushi, O.M.: Gpt-4o: the cutting-edge advancement in multimodal LLM. Authorea Preprints (2024)
14. Isotani, S., Bittencourt, I.I., Challco, G.C., Dermeval, D., Mello, R.F.: Aied unplugged: leapfrogging the digital divide to reach the underserved. In: International Conference on Artificial Intelligence in Education, pp. 772–779. Springer (2023)
15. Jeon, J., Lee, S.: Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. Educ. Inf. Technol. **28**, 15873–15892 (2023)

16. Karousos, N., Vorvilas, G., Pantazi, D., Tzagarakis, M., Karacapilidis, N., Verykios, V.S.: Summarizing responses to open-ended questions in educational surveys: a microservice-oriented approach<? show [aq id= aq1]?>. Intell. Decis. Technol. 18724981241303216 (2024)
17. Memarian, B., Doleck, T.: Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (AI), and higher education: a systematic review. Comput. Educ. Artif. Intell. 100152 (2023)
18. Monteiro Santos, M., et al.: Near feasibility, distant practicality: Empirical analysis of deploying and using LLMs on resource-constrained smartphones. In: ICTD 2024, Proceedings of the 13th International Conference on Information & Communication Technologies and Development, p. 224–235. Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3700794.3700817
19. Rodrigues, L., Pereira, F.D., Cabral, L., Gašević, D., Ramalho, G., Mello, R.F.: Assessing the quality of automatic-generated short answers using GPT-4. Comput. Educ. Artif. Intell. 100248 (2024)
20. Rodríguez, M.R.B., Cobo, C.: COVID-19 and education in the global South: emergency remote learning solutions with long-term implications (2022). https://doi.org/10.1787/897cd0d8-en, https://www.oecd-ilibrary.org/content/component/897cd0d8-en
21. Scheffé, H.: The Analysis of Variance. Wiley Series in Probability and Statistics, Wiley, New York (1959), reprinted by John Wiley & Sons in 1999
22. StatCounter: Mobile vendor market share worldwide (2024). https://gs.statcounter.com/vendor-market-share/mobile/worldwide/#monthly-202401-202501. Accessed 14 Feb 2024
23. Team, C.A.: DeepEval: open-source evaluation framework for LLMs (2023). https://github.com/confident-ai/deepeval
24. team, M.: MLC-LLM (2023). https://github.com/mlc-ai/mlc-llm
25. Tenório, K., Dermeval, D., Monteiro, M., Peixoto, A., Silva, A.P.D.: Exploring design concepts to enable teachers to monitor and adapt gamification in adaptive learning systems: a qualitative research approach. Int. J. Artif. Intell. Educ. **32**(4), 867–891 (2022)
26. Wang, B., Liu, J., Karimnazarov, J., Thompson, N.: Task supportive and personalized human-large language model interaction: a user study. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval, pp. 370–375 (2024)
27. Wang, B., Li, G., Li, Y.: Enabling conversational interaction with mobile UI using large language models. In: CHI 2023, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3580895
28. Xia, Q., Chiu, T.K., Zhou, X., Chai, C.S., Cheng, M.: Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. Comput. Educ. Artif. Intell. 100118 (2022)
29. Xiao, J., Huang, Q., Chen, X., Tian, C.: Large language model performance benchmarking on mobile platforms: a thorough evaluation. arXiv:2410.03613 (2024)
30. Yin, W., Xu, M., Li, Y., Liu, X.: LLM as a system service on mobile devices. arXiv:2403.11805 (2024)

Personalized Math Learning for Young Learners: Preliminary Findings from a Rigorous Study of a Game-Based Program

Mingyu Feng[1(✉)], Linlin Li[1], Chunwei Huang[1], Natalie Brezack[1], and Hee Jin Bang[2]

[1] WestEd, San Francisco, CA 94107, USA
{mfeng,lli,chuang,nbrezac}@wested.org
[2] Age of Learning, Inc., Glendale, CA 91203, USA
heejin.bang@aofl.com

Abstract. Building a strong foundation to support lifelong learning and success is a broadly recognized goal for early math education. Given the demonstrated benefits of personalized learning, schools and districts seek tools and approaches that engage active learners and enable personalized learning. Game-based adaptive programs provide the opportunity to achieve both goals. A rigorous large-scale study is currently undergoing to examine the efficacy of a highly engaging educational technology innovation targeting early elementary students – *My Math Academy.* Preliminary findings suggest that the program can be flexibly implemented with adequate fidelity and students' learning outcome is highly correlated with their activities and performance within the program.

Keywords: Early math learning · game-based learning · personalized · adaptive learning

1 Introduction

Developmental and cognitive theories highlight the crucial role of early mathematics skills and knowledge in predicting later academic success and preparedness for 21st-century STEM careers [1–3]. The results from the U.S. National Assessment of Education Progress (NAEP), however, show that only 39 percent of fourth grade students performed at or above the proficient level in math [4]. The situation is even more acute for children from low socioeconomic backgrounds who begin school with significantly less math knowledge [5] and demonstrate gaps in the development of early math skills relative to their middle-class peers [5, 6].

There is an urgent need to implement innovative approaches that can improve math performance. However, early learners start school with various levels of prior math knowledge and teachers face the challenge of identifying individual student needs and providing personalized instructions [7]. Well-designed, adaptive educational technology innovations offer a promising solution, enabling targeted and differentiated instruction

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 168–175, 2025.
https://doi.org/10.1007/978-3-031-99267-4_21

to help teachers better prepare all students for success in math. Digital educational games leverage play, which is important for children's cognitive development and learning [8, 9], and foster engagement and motivation by providing adaptive challenges, interactivity, and ongoing feedback [10, 11].

My Math Academy (MMA) is a game-based, highly engaging educational technology program targeting early elementary students. As a supplemental curriculum, MMA addresses math topics in alignment to the U.S. Common Core State Standards in Mathematics in early elementary grades. MMA's mastery-based learning ecosystem consists of (a) game-based activities with adaptive learning trajectories that allow students to receive differentiated support and just-in-time feedback within story contexts, (b) educator dashboards that help teachers support students' learning, and (c) offline activities that extend in-game learning experiences. The educator dashboards allow teachers to monitor individual and group progress and assign learning activities to be completed in class or at home for additional reinforcement of specific concepts (Fig. 1).

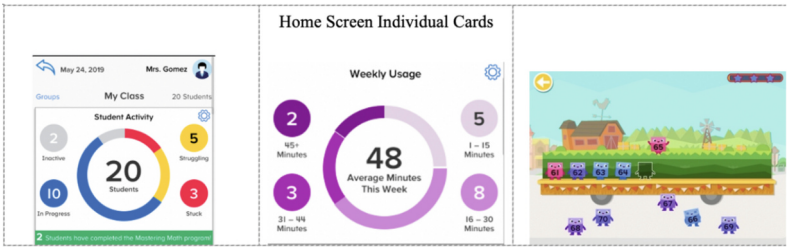


Fig. 1. Screenshots of *My Math Academy* educator dashboard (left, middle) and learner interface (right)

MMA has been shown significantly improve young students' (ages 4–6) math abilities in two pilot randomized controlled trial (RCT) studies [12, 13]. Thai and Li [12] randomly assigned 20 public school kindergarten and transitional kindergarten classrooms to either use MMA for 12 weeks or continue with business-as-usual instruction. The study found that students in the treatment group had statistically significantly higher scores on the post-test after controlling for pre-test differences ($p = 0.03$, effect size = 0.23). Another RCT study conducted in 2019 involving 22 kindergarten classrooms also showed that students who used MMA for 14 weeks significantly outperformed a comparison group of students (effect size = 0.11), and the effect was stronger for kindergarten students who began the year with significantly lower math achievement (effect size = 0.16). Other studies have shown that the majority of students who used MMA regularly ended the school year on track in math [14]. Educators regularly reported that MMA had a positive impact on students' enjoyment, interest, and self-confidence in learning math.

Supported by the U.S. Department of Education, we are conducting a nation-wide large-scale RCT[1] to provide more generalizable insights about the efficacy of MMA in improving kindergarten students' math learning. This late-breaking results paper reports on the findings from the recently completed preliminary analysis of cohort 1 of the study (2023–24 school year) and addresses the following two research questions: a) *What is the impact of MMA on kindergarten students' math learning outcome after using MMA for one school year?* b) *What is the relationship between treatment students' learning outcomes and their activities and performance within MMA?*

2 My Math Academy, the Personalized Learning Program

MMA is a game-based, adaptive learning system designed to help students build a strong understanding of fundamental number sense and operations. Topics range from counting to ten to adding and subtracting three-digit numbers using the standard algorithm. MMA includes 31 games with over 300 game-based activities, covering 130 number sense and operations concepts and skills for prekindergarten through second grade, aligning with state and national standards (e.g., Common Core State Standards in Mathematics, National Council of Teachers of Mathematics' Principles Standards for School Mathematics). It is a Digital Promise and STEMworks-certified program and has been used by over 150,000 teachers each year in the U.S. since 2018.

As a personalized mastery-based learning ecosystem, MMA leverages game-based learning [16] and uses evidence-centered design [16] to enable learners to master math concepts through playful experiences. MMA supports students' independent practices and connects real-world experiences through *game-based* learning activities within *story contexts*. The story context of a math problem makes concepts and operations more meaningful to students and provides students with a framework for understanding what they are expected to do and why [17]. Storylines in game-based learning activities help all students, including struggling readers, gain access to the math problems, make sense of math problems, and transfer skills learned in games into the real world. The learning activities also enable integrated, ongoing formative assessment to provide immediate, just-in-time *feedback* to students during the learning process, which enables ongoing feedback cycles and customized learner difficulty levels [18].

MMA offers adaptive learning activities and allows students to receive personalized, differentiated support. The adaptive games provide dynamic feedback to guide students; if students make mistakes, they receive increasing levels of scaffolding based on their performance. Teachers can view the educator dashboard, which provides real-time student progress data about the amount of time students have been engaged in MMA, their in-game performance, and the learning objectives on which students are struggling or excelling. The dashboard also suggests offline activities for students stuck on certain learning objectives, allowing teachers to better tailor their instruction to accelerate learning [19]. Driving MMA's adaptive engine is a knowledge map, a blueprint of granular,

[1] The study has been approved by the Institutional Review Board at WestEd. Participating teachers all signed consent forms. Parents received opt-out or consent forms to give permission for their children to participate in the study. The study was pre-registered at the Registry of Efficacy and Effectiveness Studies, #15920.1v1.

measurable learning objectives and paths to the development of early number sense. Activities that are under the same math skill are grouped into nodes. Nodes often have a progression of activities from teaching new concepts or skills, through practice, into mastery. The nodes that share a content theme are grouped into modules. Figure 2 shows an example of how a node map is structured. Given these adaptivity and scaffolding mechanisms at multiple levels, each student has a completely personalized experience, tailored precisely to their "ready to learn" math level and learning pace.

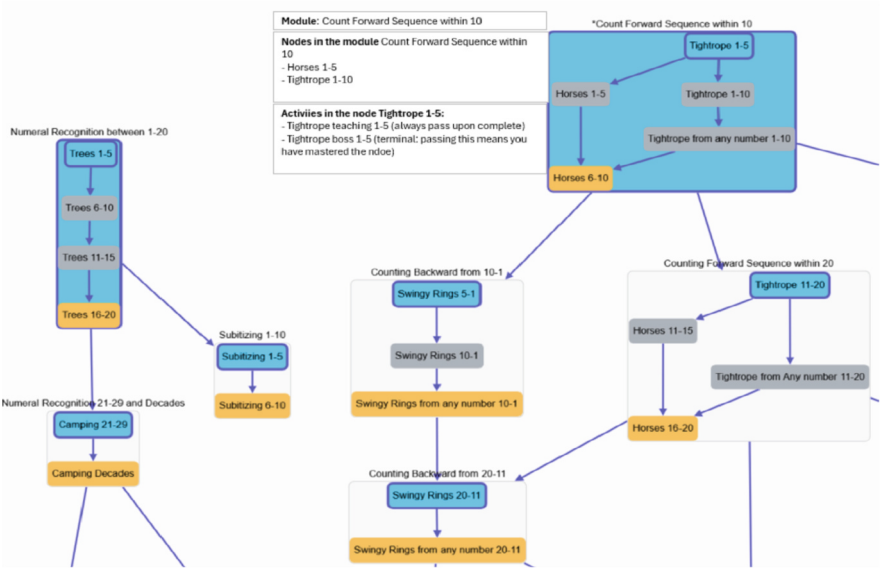


Fig. 2. A Sample of the Node Map of Games in *My Math Academy*

3 Research Design

The RCT used a multi-site, school-level, clustered randomized experimental design. Schools within each district were paired based on their demographic characteristics and then randomly assigned to a treatment or business-as-usual control condition. All consented kindergarten teachers in a school were assigned to the same condition. The RCT was conducted by an organization independent of MMA's developers.

We have successfully recruited and randomized 44 elementary schools across four states from West, Central, and Southwestern of the United States (14 schools in Cohort 1 and 30 schools in Cohort 2). Cohort 1 included 14 schools (36 kindergarten teachers and their 753 students, ages 5–7) that were distributed across rural, suburban, and city communities in five public school districts. Seven schools with 21 classrooms ($n = 429$ students) were randomly assigned to the treatment group where teachers received training and implemented MMA as a supplemental program for classroom math learning support over the entire school year of 2023–24. Seven schools with 15 classrooms (n

= 324 students) were randomly assigned to the control group where teachers continued their existing instructional practices, including any use of online programs, but not had access to MMA.

4 Data, Analysis and Findings

4.1 Data Collection

Students' learning outcome was measured by the Test of Early Mathematics Ability-3 (TEMA; [20]), a standardized, nationally normed achievement test measuring thee- to eight-year-old children's informal and formal math knowledge. Trained and certified assessors conducted one-on-one pre- and post-assessments with a subsample of 367 cohort 1 students individually in-person to ensure reliable data collection. Student scores from the TEMA assessment were used as a baseline measure in fall 2023 and as an outcome measure in spring 2024. Individual **student demographics** including race, ethnicity, gender, age, and socioeconomic status were collected from school districts.

MMA platform usage data was used to monitor students' actual exposure to MMA. MMA collects real-time backend electronic records of students' game play data, including all student actions and responses as well as their progress on grade-level skills. Metrics were built upon the records to track student in-system performance, usage (time spent, activities completed), progress through the system, and performance on each activity. **Teacher interview, logs, and classroom observations** were conducted to capture day-to-day use and any challenges to implementation. Teachers provide their perspectives about the effectiveness of MMA (or other technology-based programs for the control condition) and their student engagement when using the program. Qualitative analysis and descriptive analysis provide important implementation context and help with the interpretation of the impact results.

4.2 Analysis and Results

To evaluate the efficacy of MMA for improving students' math learning in kindergarten (RQa), we conduct an intent-to-treat (ITT) analysis of cohort 1 student outcome data, using a two-level hierarchical linear regression model (HLM) to account for students nested within schools. We modeled the mean differences in TEMA post-test scores between students in treatment and control schools, controlling for student's pre-test scores and other student- and school-level covariates to improve the precision of the impact estimate. To explore the relationships between the treatment students' MMA usage, within-game performance, and math achievement (RQb), we conducted descriptive and correlation analyses of MMA usage metrics and students' pre- and post- TEMA assessment scores. To better understand implementation, teacher interviews and observations were analyzed qualitatively using thematic analysis [21] and constant comparison [22], and teacher logs were analyzed descriptively.

The preliminary result based on cohort 1 data indicates that kindergarten students in the treatment group generally outperformed those in the control group on the post-TEMA assessment, after accounting for baseline assessment scores. However, the difference was

not significant at the 0.05 level. Given the limited sample size of 14 schools, the findings should be interpreted with caution.

Analysis of MMA usage data suggests that the program can be flexibly implemented with adequate fidelity. Treatment teachers integrated MMA into their instruction for 20–26 weeks during the school year, with students actively engaged in in-game activities for at least 25 min per week. Correlational analyses revealed strong positive relationships ($r > 0.44$) between engagement metrics—such as total time spent, average weekly minutes, and active weeks—and in-game learning outcomes. Furthermore, students who demonstrated greater improvement on the TEMA assessment tended to complete and pass more activities within MMA, mastering more skills at or below their grade level. Notably, students who practiced more on number relations and number operations topics within MMA showed greater gains on the TEMA assessment. While the analysis was correlational, these findings support the validity of in-program metrics as continuous measures of student math performance and highlight the potential of MMA to enhance student learning when implemented with adequate dosage.

Qualitative analysis of cohort 1 teacher interviews and logs suggests that MMA can be flexibly implemented either with the whole class or in small groups. In general, findings indicated that students were engaged while using the program, but engagement varied over the school year. Most teachers felt that MMA supported their students' math learning, particularly by facilitating differentiated learning – MMA offered targeted support for struggling learners and enrichment opportunities for more advanced students. For example, one teacher shared that compared to prior school years, with MMA, students were learning math more quickly: *"I feel like several of them have gotten more confident [and] picked up the counting a lot quicker."*

5 Conclusion and Future Work

MMA illustrates how effective design strategies can empower technology-based programs by incorporating playful games, contextualized instruction, and personalized learning pathways to support young children's engagement in math learning. The preliminary results suggest that MMA can be flexibly and regularly integrated into instruction and shows potential for engaging students and improving learning of grade-level content. It shines light on how to implement an in-depth, technology- and game-based program to empower teachers, enable customized learning opportunities to address individual student needs, and build an engaging environment for all students in math.

Currently, implementation and data collection are underway in 30 cohort 2 schools during the 2024–25 school year. Future analysis will include the full sample of the large-scale study to discuss the impact of MMA on student math learning, addressing whether the effects of MMA vary for students with different demographic characteristics or different prior achievement levels, and further investigate factors affecting program implementation.

Acknowledgments. The research is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A220159 awarded to WestEd. 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. *My Math Academy* pre-K through 5th grade is a product owned by Age of Learning. The efficacy study, including data collection and analysis, is conducted independently by WestEd. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chu, F.W., vanMarle, K., Geary, D.C.: Predicting children's reading and mathematics achievement from early quantitative knowledge and domain-general cognitive abilities. Front. Psychol. **7**, 1–14 (2016)
2. Watts, T.W., Duncan, G.J., Siegler, R.S., Davis-Kean, P.E.: What's past is prologue: relations between early mathematics knowledge and high school achievement. Educ. Res. **43**(7), 352–360 (2014)
3. Devlin, B.L., Jordan, N.C., Klein, A.: Predicting mathematics achievement from subdomains of early number competence: differences by grade and achievement level. J. Exp. Child Psychol. **217**, 105354 (2022)
4. National Center for Education Statistics (NCES): The nation's report card (2024). https://www.nationsreportcard.gov/reports/mathematics/2024/g4_8/national-trends/?grade=4
5. Nores, M., Barnett, S.: Access to high-quality early care and education: readiness and opportunity gaps in America. Center for Enhancing Early Learning Outcomes and National Institute for Early Education Research, New Brunswick, NJ (2014)
6. Starkey, P., Klein, A., Wakeley, A.: Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. Early Child. Res. Q. **19**(1), 99–120 (2004)
7. Goddard, Y., Goddard, R., Kim, M.: School instructional climate and student achievement: an examination of group norms for differentiated instruction. Am. J. Educ. **122**(1), 111–131 (2015)
8. Piaget, J.: The stages of the intellectual development of the child. Bull. Menninger Clin. **26**(3), 120 (1962)
9. Fisher, K., Hirsh-Pasek, K., Golinkoff, R.M.: Fostering mathematical thinking through playful learning. In: Suggate, S., Reese, E. (eds.) Contemporary Debates on Child Development and Education, pp. 81–92. Routledge, London (2012)
10. Gee, J.P.: Good Video Games and Good Learning: Collected Essays on Video Games, Learning, and Literacy. Peter Lang Inc., New York (2007)
11. Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. **78**(1), 153–189 (2008)
12. Thai, K., Bang, H.J., Li, L.: Accelerating early math learning with research-based personalized learning games: a cluster randomized controlled trial. J. Res. Educ. Effect. **15**, 28–51 (2022)
13. Bang, H.J., Li, L., Flynn, K.: Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning. Early Child. Educ. J. **51**, 717–732 (2023)
14. Bang, H & Thai, K.: Prekindergarten students using My Math Academy significantly improved their school readiness in math during a school year disrupted by the pandemic. Age of Learning (2022). https://www.ageoflearning.com/wp-content/uploads/2023/09/HARLINGEN_MM_ResrchBrf_RGB_12.pdf
15. Plass, J.L., Homer, B., Kinzer, C.: Foundations of game-based learning. Educ. Psychol. **50**, 258–283 (2016)
16. Mislevy, R.J., Almond, R.G., Lukas, J.F.: A brief introduction to evidence-centered design, p. 37. Educational Testing Service, Princeton, NJ (2003). https://www.ets.org/Media/Research/pdf/RR-03-16.pdf

17. Sullivan, P., Zevenbergen, R., Mousley, J.: The contexts of mathematics tasks and the context of the classroom: are we including all students? Math. Educ. Res. J. **15**(2), 107–121 (2003)
18. Shute, V.J., Kim, Y.J.: Formative and stealth assessment. In: Spector, J.M., Merrill, M.D., Elen, J., Bishop, M.J. (eds.) Handbook of Research on Educational Communications and Technology, pp. 311–321. Springer, New York, NY (2014)
19. Gersten, R., et al.: Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools. NCEE 2009–4060. What Works Clearinghouse (2009)
20. Ginsburg, H.P., & Baroody, A.J.: Test of early mathematics ability third edition examiner's manual. Pro-ed, Austin, TX (2003)
21. Onwuegbuzie, A.J., Teddlie, C.: A framework for analyzing data in mixed methods research. In: Tashakkaori, A. and Teddlie, C. (eds.) Handbook of Mixed Methods in Social & Behavioral Research, pp. 397–430. SAGE (2003)
22. Strauss, A., Corbin, J.: Basics of Qualitative Research: Grounded Theory Procedures and Techniques, 3rd edn. Sage, Newbury Park (2008)



Teachers' Perspectives on Using and Teaching Artificial Intelligence in Early Primary Education

Marius Vogt[1(✉)], Valentina Ferraioli[2], Victoria Abou-Khalil[2], Lena Hollenstein[1], Francesco Mondada[2], and Franziska Vogt[1]

[1] St. Gallen University of Teacher Education, 9000 St. Gallen, Switzerland
marius.vogt@phsg.ch
[2] Swiss Federal Institute of Technology Lausanne (EPFL), 1010 Lausanne, Switzerland

Abstract. As artificial intelligence (AI) becomes increasingly integrated into various aspects of daily life, the need for AI literacy continues to grow. Accordingly, both educators and students must develop a critical understanding of AI systems and acquire the skills to apply them responsibly. However, AI literacy education is associated with barriers for teachers. This study examines teacher perspectives on AI literacy education through focus group discussions and single interviews with 18 primary school teachers from all of Switzerland (French- Italian- and German-speaking regions). For the analysis of the interviews, a summarizing qualitative content analysis was applied, incorporating barriers to technology integration and AI literacy education taken from the research literature. The analysis sheds light on barriers to AI literacy education in the context of early primary education. The results indicate that barriers identified in the literature were also present in the context of primary school teachers in Switzerland. Moreover, it became apparent that misconceptions about the scope of AI are closely linked to a lack of ideas for pedagogical implementation of AI literacy. Additionally, teachers' concerns about AI's high transformative potential might be linked with teachers' hesitation to integrate AI literacy education in their classroom. Implications for teacher professional development (PD) are derived.

Keywords: AI literacy · primary school · teacher interviews · barriers

1 Introduction

The ongoing digitalisation is accompanied by novel competency requirements for students and teachers. A salient aspect of digitalisation is the increasing prevalence of AI in various domains of everyday life [1]. Consequently, AI literacy is being recognised as a key skillset for all individuals, including young children [2–5]. Despite this recognition, in general, AI literacy is not yet integrated in education. Specifically in the context of early primary education, implementation is met with concern, as screen-based activities for engaging with AI are not suitable for children as they are still in the process of learning to read and write. In order to effectively support teachers in developing their

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 176–183, 2025.
https://doi.org/10.1007/978-3-031-99267-4_22

own AI literacy and providing AI literacy education in the classroom, it is essential to first understand teachers challenges and needs. Therefore, the objective of this study is to capture the perspectives of primary school teachers from the first two grades in three language regions of Switzerland on using and teaching AI. The research question is: What are teachers' perspectives in general to AI literacy education in first two grades of primary school in Switzerland and what are the specific barriers to AI literacy education coming to the fore?

2 Theory and Related Work

There is consensus that AI literacy is essential in the 21st century and should be fostered from an early child age on [2–5]. AI literacy includes the abilities to understand, use, critically evaluate and ethically engage with AI technologies [4]. For early primary education it can be suggested that AI literacy encompasses a basic conceptual understanding of how AI systems learn and generate outputs, basic applications of AI technologies and the cultivation of a responsible human-centered perspective. Despite the acknowledged importance and potential of AI literacy, challenges and barriers persist for teachers in providing effective learning opportunities. This applies particularly to young children who have not yet developed reading and writing skills. To design effective and continuous PD for teachers of young children, it is essential to understand the challenges and barriers they face. Barriers to technology integration can be categorized within the framework proposed by Hew and Brush [6]. Building on Ertmer's [7] typology of external and internal barriers, this framework identifies six key themes as technology integration barriers, namely as external barriers resources, institution, subject culture, and assessment and as internal barriers attitudes and beliefs and knowledge and skills. Although the review dates back to 2007, it remains a relevant framework for systematizing barriers. More recent studies in fields such as computer science [8], ICT [9–11], and digital citizenship [12–14] continue to identify barriers that align with Hew and Brush's [6] framework, thereby underscoring its continued applicability.

In the educational context, the following is known about the barriers to ensuring children's AI literacy education. At the level of external barriers, a lack of curricular guidelines and institutional support stand out as barriers in early childhood education [5, 15] and in primary school [16]. Lack of resources is perceived as a barrier, with issues such as inadequate technological infrastructure and budget constraints [17, 18]. Moreover, scepticism towards AI among stakeholders, ranging from school leadership to parents, can further impede efforts to implement AI literacy education [17, 18]. Ethical concerns, along with the lack of clear policies on AI use in schools, present an additional barrier to the integration of AI into the educational system [17].

An internal barrier is teachers' lack of knowledge, skills, confidence, and pedagogical preparedness for effectively integrating AI literacy education [5, 15, 18]. Closely linked to this is a deficiency in teachers' technological pedagogical content knowledge (TPACK), which affects teachers' confidence in both adopting AI tools for teaching [17] and providing learning opportunities to develop AI literacy [18]. Another barrier is teachers' perception of AI as being overly complex for young children [15]. Furthermore, concerns were raised about the rapid pace of AI development making educational

content quickly outdated [16]. In addition, generally negative attitudes towards AI can present a barrier [5, 16, 18].

The studies show that there are various external and internal barriers to implementing learning opportunities to develop AI literacy in young children. As it is crucial to equip children with knowledge and skills regarding AI, teachers need to develop professional skills for teaching AI literacy education to children in the early years of primary school. In order to take teachers' perspective into account and to understand the barriers hindering implementation of AI literacy education in the first years of primary school, more research is required. To contribute to this research gap, the current study takes a qualitative in-depth look at AI literacy barriers perceived by primary school teachers in Switzerland.

3 Methods

To capture primary school teachers' perspective on AI education, qualitative semi-structured focus group and single teacher interviews were conducted. For this purpose, $N = 18$ primary school teachers (14 female and 4 male teachers) from the Italian, French and German speaking part of Switzerland were recruited via convenience sampling. The inclusion criterion was employment as a teacher in one of the first two grades of primary school. The interviews were conducted between June and October 2024. No additional background variables were collected. Ethical approval was obtained from the ethics board of EPFL.

In order to structure the interview process, a semi-structured interview guide was developed in advance. Key guiding questions included "In your impression, what is the current relevance of AI-related skills for children in the first to years of primary school?" and "What relevance do you envisage for AI-related skills for children in the first two years of primary school?". Similar questions were asked regarding the competencies of teachers: "In your impression, what is the current relevance of AI-related skills for primary school teachers?" and "What relevance do you envisage for primary school teachers?". The interviews were recorded with the consent of the participants. In total, audio and video recordings were made with an overall duration of 242 min. The recordings were transcribed using MAXQDA software, following the transcription guidelines of Dresing and Pehl [19]. The data were subsequently anonymized to ensure participant anonymity.

The transcripts were analysed in accordance with the summarising and structuring qualitative content analysis outlined by Mayring [20], combining an inductive and deductive analysis. In the first round of the analysis, categories were developed inductively from the text. In the second round, these categories were abstracted and subsumed. At this stage, twenty percent of the German-language transcripts were double-coded, achieving a satisfactory intercoder agreement of $\kappa = 0.71$ [21]. For the coding of the Italian and French transcripts, a consensual agreement approach was used, double coding all data and discussing categorization. This method ensured that potential language barriers in double coding by non-native speakers were effectively mitigated. In the third round, the barrier categories derived of the research literature were applied to the material in a deductive approach [20].

4 Results

The result of the content analysis was a system of deductive and inductive barrier categories (see Table 1).

Resources
In terms of resources, teachers highlighted time constraints both for developing their own AI literacy and for creating learning opportunities that foster AI literacy in children. Additionally, the analysis revealed that the availability of suitable technological tools and devices varies to a very large extent: whereas some schools have access to suitable devices, to other schools, the lack of resources is a crucial barrier.

Institution
Curriculum Design and *Ethical Guidelines.* Teachers mentioned the absence of clear curricular and ethical guidelines on how AI should be responsibly implemented in the classroom and which tools should be used.

Moreover, teachers allocate the responsibility of implementing AI literacy education to their school principals and are sceptical of individual teacher-driven solutions. Simultaneously, they reported that the emphasis on digital education and ensuring AI literacy at their respective schools depends largely on the personal interests and attitudes of the principals. Aspects related to *subject culture* and *assessment practices* were not mentioned in any of the teachers' statements.

Attitudes and Beliefs
Importance of AI. Part of the teachers acknowledged the growing significance of AI in society and recognized the relevance of AI literacy for both teachers and students. These teachers expressed openness towards integrating AI into the curriculum and they demonstrated an interest in expanding their own understanding of AI. However, a contrasting perspective also emerges within the sample. Some teachers do not perceive AI as an urgent topic and therefore see no necessity to focus on developing AI literacy in primary schools. In general, teachers do not consider an in-depth understanding of AI as necessary. Even teachers who consider it important to introduce AI in primary school do not view it as a central topic. Instead, they perceive it as one of many competencies that children need to acquire at this stage, and it was pointed out that children should not be overburdened. AI literacy education is seen as particularly valuable when it also supports the development of transversal skills, such as critical thinking.

Children's Comprehensive Ability. Teachers perceive the topic as complex and are uncertain whether children's cognitive abilities are sufficient to grasp AI concepts.

Transformative Potential. The interviews revealed a perception among teachers of AI as a powerful, fast-moving force. They perceive AI as capable of driving lasting changes in both society and education. Scenarios in which AI takes over an increasing number of tasks are regarded as realistic, leading to uncertainty in assessing the competencies still required for both teachers and students.

Table 1. Category system of AI literacy barriers.

Category I	Category II	Example
External	Resources [6, 15, 17, 18]	"Because otherwise it simply takes a lot of time if you want to go into it (AI) in depth"
	Institution [5, 6, 15, 17, 18]	"If a school principal is interested, it's great, and if not, it (AI PD) is simply not offered"
	Curriculum design [5, 15, 16]	"That it is not simply, at primary school we deal with it (AI) like this and then at secondary school like that. So, I think there should be a common thread"
	Ethical guidelines [17]	"I don't know who is behind it (AI) or where this data is stored"
	Subject culture [6]	
	Assessment practices [6]	
Internal	Attitudes and beliefs	
	Importance of AI [5, 6, 16, 18]	"It might not be so bad if you know a bit, but I don't think you have to delve too deeply into it"
	Children's comprehensive ability [15]	"I think the whole thing (AI) is really still a bit too complex for the children"
	Transforming potential	"I think something big is coming our way. Not just for the school, but also for myself"
	Risks of AI	"That self-efficacy is also lost a little through this. Feeling able to do it yourself"
	Knowledge and skills	
	Content knowledge [5, 6, 15–18]	"I myself am still quite uneducated in this area"
	Pedagogical know-ledge [6, 16–18]	"I cannot imagine that I could somehow incorporate this (AI) into the lessons for the children right now"
	Parent Communication [15]	"And I wouldn't know where to draw the line, what to tell the children, what not to tell the children and certainly not the parents"

Risks of AI. Teachers identify various risks associated with the implementation of AI in schools. These include dependency and over-reliance on AI that could undermine children's effective learning strategies and critical thinking. Additionally, some teachers

express concerns about a loss of autonomy in teaching, extending to the possibility of being replaced by AI. Furthermore, excessive media and internet consumption is viewed as harmful, with AI being perceived as part of this issue.

Knowledge and Skills

Content Knowledge. Most teachers express a lack of competence, confidence and experience when it comes to understanding, using and teaching AI. Consequently, teachers express reluctance using AI and implementing AI literacy. It was also evident that a misconception exists among teachers. Teachers equate AI with Large Language Models, specifically ChatGPT. Rare descriptions of examples for AI classroom implementation or personal use exclusively referred to ChatGPT or, in one case, image-generating AI.

Pedagogical Knowledge. When selecting instructional content, teachers assign a central role to specially trained teachers, referred to as Pedagogical ICT Support (PICTS). These specialists prepare relevant content for them and provide guidance on what and how to implement in practice. Indeed, for most teachers, it is challenging to envision age-appropriate ways to help young primary school children develop AI literacy.

Parent Communication. In addition, teachers are unsure how to communicate with parents about AI. This is particularly important, as teachers perceive that inequalities in AI literacy emerge from an early age, primarily influenced by the home environment.

5 Conclusion and Discussion

The study reported here is based on interviews with 18 primary school teachers, which were analysed through qualitative content analysis with respect to barriers teachers perceive to the implementation of AI literacy education. The resulting category system revealed that barriers identified in the literature [5, 15–18] also emerge in the context of Swiss primary schools. These barriers align with the framework proposed by Hew and Brush [6]. The absence of *Assessment practices* as a barrier is likely due to the fact that grading does not play a significant role in the first two years of primary school in Switzerland. The lack of *Subject culture* as a perceived barrier can be attributed to primary school teachers being generalists rather than being trained to teach specific subjects. The interview analysis uncovered an expansion of the barriers, providing a more differentiated view. As an inductive category at the level of attitudes and beliefs, the interviews revealed a belief in the transformative power of AI, accompanied by concerns about potential risks for education. At the knowledge level, a lack of knowledge, but also misconceptions became apparent. Moreover, the in-depth analysis of the interviews revealed how interconnections between various barriers emerge. For example, the misconception of equating AI with ChatGPT gives rise to other barriers. Since ChatGPT requires advanced literacy skills, teachers perceive AI literacy (understood in relation to ChatGPT) as too complex for the early years of primary school. Consequently, teachers struggle to envision targeted learning activities for young children.

Furthermore, the interviews revealed that teachers navigate a tension between caution and curiosity. On one hand, they express concerns about potential over-reliance and

dependency on powerful and unpredictably evolving AI. On the other hand, they recognize the importance of AI literacy as a 21st century skill. To successfully manage this balancing act, they consider themselves dependent on support from experts and school leaders, emphasizing the crucial role of institutional backing. Moreover, most teachers express openness to PD opportunities for developing AI literacy and implementing learning opportunities for children to acquire AI literacy.

There are aspects which are important to keep in mind when interpreting the results of this study: The sample is not representative, and therefore, the findings cannot be generalized to all primary school teachers in Switzerland. Nevertheless, efforts were made to ensure diverse perspectives in the sampling process by including teachers from various cantons across three language regions. The questions posed to teachers were formulated in an open-ended manner without introducing a specific concept of AI literacy. This approach had the advantage of capturing teachers' conceptual understanding of AI more accurately. However, it also made it more difficult to identify nuanced positions on AI literacy education.

Based on the identified barriers, the following implications could be drawn for teacher PD: The PD should provide conceptual knowledge about AI that goes beyond the mere application of ChatGPT. For example, it can be shown how AI concepts are an extension to Computational Thinking skills described in existing curricula. This would make entry points for the implementation of AI in primary schools more apparent to teachers. Additionally, the PD should introduce age-appropriate activities for AI literacy education in early primary school. Promising approaches to promote AI literacy in early childhood are activities with learnable machines and robots [5]. Moreover, screenless activities that do not require reading and writing skills such as social pretend play [22] could offer accessible entry points to AI literacy. Furthermore, it is essential that PD addresses teachers' concerns regarding the transformative potential and risks of AI. PD should equip teachers with the ability to identify responsible ways to benefit from AI and to be aware of its limitations.

Acknowledgments. We would like to thank all the teachers for participating in the interviews and for the helpful insights they provided. This work was supported by the Jacobs Foundation as part of the consortium DEEP (Digital Education for Equity in Primary Schools).

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

References

1. Ng, D.T.K., Leung, J., Su, J., Yim, I., Qiao, M., Chu, S.: AI Literacy in K-16 Classrooms. Springer, Cham (2022)
2. Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., Huber, P.: Artificial intelligence and computer science in education: from kindergarten to university. In: 2016 IEEE Frontiers in Education Conference (FIE), pp. 1–9 (2016)
3. Miao, F., Shiohira, K.: AI Competency Framework for Students. UNESCO, Paris (2024)
4. Ng, D.T.K., Leung, J.K.L., Chu, K.W.S., Qiao, M.S.: AI literacy: definition, teaching, evaluation and ethical issues. Proc. Assoc. Inf. Sci. Technol. **58**(1), 504–509 (2021)

5. Su, J., Ng, D.T.K., Chu, S.K.W.: Artificial Intelligence (AI) literacy in early childhood education: the challenges and opportunities. Comput. Educ.: Artif. Intell. **4**, 100124 (2023)
6. Hew, K.F., Brush, T.: Integrating technology into K-12 teaching and learning: current knowledge gaps and recommendations for future research. Educ. Tech. Res. Dev. **55**(3), 223–252 (2007)
7. Ertmer, P.A.: Addressing first-and second-order barriers to change: strategies for technology integration. Educ. Tech. Res. Dev. **47**(4), 47–61 (1999)
8. Yadav, A., Gretter, S., Harmbrusch, S., Sands, P.: Expanding computer science education in schools: understanding teacher experiences and challenges. Comput. Sci. Educ. **26**(4), 235–254 (2016)
9. Heitink, M., Voogt, J., Verplanken, L., van Braak, J., Fisser, P.: Teachers' professional reasoning about their pedagogical use of technology. Comput. Educ. **101**, 70–83 (2016)
10. Niederhauser, D.S., et al.: Sustainability and scalability in educational technology initiatives: research-informed practice. Tech Know Learn. **23**(3), 507–523 (2018)
11. Passey, D.: Computer science (CS) in the compulsory education curriculum: Implications for future research. Educ. Inf. Technol. **22**(2), 421–443 (2017)
12. Choi, M., Cristol, D., Gimbert, B.: Teachers as digital citizens: the influence of individual backgrounds, internet use and psychological characteristics on teachers' levels of digital citizenship. Comput. Educ. **121**, 143–161 (2018)
13. Martin, F., Gezer, T., Wang, W.C., Petty, T., Wang, C.: Examining K-12 educator experiences from digital citizenship professional development. J. Res. Technol. Educ. **54**(1), 143–160 (2022)
14. Walters, M.G., Gee, D., Mohammed, S.: A literature review: digital citizenship and the elementary educator. Int. J. Technol. Educ. **2**(1), 1–21 (2019)
15. Su, J.: Kindergarten teachers' perceptions of AI literacy education for young children. Int. J. Technol. Des. Educ. **34**(5), 1665–1685 (2024)
16. Wang, T., Cheng, E.C.K.: An investigation of barriers to Hong Kong K-12 schools incorporating Artificial Intelligence in education. Comput. Educ.: Artif. Intell. **2**, 100031 (2021)
17. Cheng, E.C.K., Wang, T.: Leading digital transformation and eliminating barriers for teachers to incorporate artificial intelligence in basic education in Hong Kong. Comput. Educ.: Artif. Intell. **5**, 100171 (2023)
18. Yim, I.H.Y., Wegerif, R.: Teachers' perceptions, attitudes, and acceptance of artificial intelligence (AI) educational learning tools: an exploratory study on AI literacy for young students. Future Educ. Res. **2**(4), 318–345 (2024)
19. Dresing, T., Pehl, T.: Praxisbuch Transkription. Regelsysteme, Software und praktische Anleitungen für qualitative ForscherInnen. Eigenverlag, Marburg (2011)
20. Mayring, P.: Qualitative Inhaltsanalyse. Beltz, Weinheim (2015)
21. Brennan, R.L., Prediger, D.J.: Coefficient kappa: some uses, misuses, and alternatives. Educ. Psychol. Measur. **41**(3), 687–699 (1981)
22. Vogt, F., Hollenstein, L.: Exploring digital transformation through pretend play in kindergarten. Br. J. Edu. Technol. **52**(6), 2130–2144 (2021)

Automatic Assessment of Spoken English for Uruguayan Spanish L1 Young Learners

Ana Clara Nóbile[1,2(✉)], Ignacio Viscardi[2], Pablo Cancela[2], Germán Capdehourat[1,2], Luis Chiruzzo[2], and Santiago Góngora[2]

[1] Ceibal, Montevideo, Uruguay
{anobile,gcapdehourat}@ceibal.edu.uy
[2] Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay
{ignacio.viscardi,pcancela,luischir,sgongora}@fing.edu.uy

Abstract. Assessing students' spoken language skills at scale remains a challenge in education. Speaking assessments typically rely on human raters, making automation a crucial step toward broader scalability. This work presents a system for automatic spoken language assessment, developed in the context of *Ceibal en Inglés*, a national educational program in Uruguay that provides English instruction to thousands of students in public primary and secondary schools. Designed to evaluate both speech production and semantic content in students' responses, the system incorporates analysis components tailored to different aspects of oral performance. The speaking assessment includes exercises where students record responses to prompts, such as short questions, extended responses, and image-based tasks. For speech production, we employ a multi-task pronunciation assessment model designed to evaluate multiple aspects of spoken output. For semantic content, we developed a pipeline starting with automatic speech recognition (ASR), followed by a large language model (LLM) guided by customized prompt sets for each test part.

We evaluated the system components using different datasets that reflect the linguistic and demographic profiles of the target student population.

The proposed approach, which combines speech analysis and Natural Language Processing, has proven to be an effective solution for addressing the automatic evaluation of spoken language and supports ongoing efforts toward its real-world implementation in large-scale educational programs.

Keywords: Automatic Speech Assessment · Large-Scale Oral Assessment · Natural Language Processing · Large Language Models

1 Introduction

Automatic assessment of spoken language is a key component of computer-assisted language learning (CALL) [1]. Early CALL systems primarily focused on

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 184–191, 2025.
https://doi.org/10.1007/978-3-031-99267-4_23

repetitive grammar and vocabulary exercises, later evolving to include interactive activities such as simulations and games [2]. More recently, advances in speech and language technologies have enabled both commercial and research-based applications to evaluate pronunciation, fluency, and even aspects of content. However, these systems often rely on proprietary methods, and no standardized framework for automated speaking assessment has been widely adopted [3].

In this context, our work focuses on evaluating oral skills within *Ceibal en Inglés*, a large-scale national educational program in Uruguay dedicated to English language learning [4]. The program serves thousands of Spanish L1 students in primary and secondary schools. It recently introduced an online speaking test in which students record responses to various prompts, ranging from short questions to guided tasks and image-based scenarios [5]. While this format enables remote data collection, manual grading remains a bottleneck, as each response requires evaluation by human raters, limiting the scalability of the assessment process.

To address this challenge, we develop and evaluate automatic methods to assess this speaking test. Specifically, we investigate how Large Language Models and recent advances in speech processing can approximate human judgments regarding two key aspects of spoken responses: their speech production characteristics and the content and coherence of the answers. Since the learners are children and teenagers whose first language is Uruguayan Spanish, we prioritize evaluation on data that reflects their accents and speaking patterns.

2 Datasets

This study primarily relies on two datasets that align with our target population.

The first one was constructed for this work based on the speaking test previously introduced and is described in Sect. 2.1.

The second, *EpaDB* [6], is a dataset of English phrases read aloud by speakers from Argentina. Originally developed for pronunciation assessment, we use a subset of it to select the ASR model for our pipeline. To the best of our knowledge, it is the only publicly available dataset with a linguistic background similar to our target population, as both are native speakers of Rioplatense Spanish.

2.1 Uruguayan Speaking Test Dataset (UST Dataset)

The *UST Dataset* comprises recordings from **23 students**, each responding to **13 different tasks**, which are distributed across three distinct parts of the test:

Part 1 Question & Answer involves straightforward questions such as "Can you swim?" or "What is your name?", designed to evaluate the student's ability to understand and respond to basic inquiries.

Part 2 Name It provides a question along with an associated image, as in "What food is it?", thereby assessing the student's capacity to interpret visual information and respond accurately.

Part 3 Talk About Familiar Topics requires students to produce extended and structured responses, such as a prompt like "Tell me about your holidays", accompanied by two guiding questions ("Where did you go?" and "What did you do?").

Assessment follows an **analytic scoring approach**, where scores are assigned **independently to each category** on a scale from **0 to 3**. These scores reflect the student's overall performance across all responses, with each one evaluated separately according to its corresponding category, without assigning labels to individual recordings. Among the categories evaluated, this study focuses on:

- **Task Fulfilment**, which assesses the overall relevance and completeness of the responses.
- **Delivery**, which evaluates the overall intelligibility and flow of speech.

To ensure data quality, we applied preprocessing steps to exclude recordings that were empty, excessively noisy, or entirely in Spanish—issues commonly encountered in educational settings.

This dataset is used for *Delivery* analysis and also serves as the foundation for constructing a separate dataset designed to evaluate response-level *Task Fulfilment*, augmented with synthetic data. Further details are in Sect. 4.

3 Delivery Evaluation

Evaluating spontaneous speech poses additional challenges compared to read speech due to its inherent variability and unpredictability [7].

To automatically assess *Delivery*, we employed *MultiPA* [8], a state-of-the-art pronunciation assessment model chosen for its open-source availability and applicability to spontaneous speech. This model was trained and evaluated on datasets consisting of Mandarin L1 speakers, which does not necessarily guarantee good performance for Uruguayan Spanish L1 learners of English. Therefore, evaluating the model on this new domain is essential for assessing its generalization capabilities and robustness to domain shifts.

This model generates four metrics (*accuracy*, *fluency*, *prosody*, and *total*) for each audio file at utterance level. Since the human-assigned *Delivery* scores are provided at the speaker level, based on the overall performance across all responses, we aggregate these metrics per speaker to enable a direct comparison between the system's predictions and the reference evaluations. Among the aggregation methods tested, the mean proved to be the most suitable.

To examine the relationship between speaker-level aggregated scores and human evaluations, we computed Spearman[1] correlation coefficients to quantify their alignment. Table 1 reports these coefficients, showing strong positive correlations across the evaluated dimensions, with all values being statistically significant ($p < 0.01$).

[1] Spearman was chosen over Pearson as it does not assume a linear relationship between model predictions and human ratings.

Table 1. Spearman correlations between *MultiPA*'s aggregated predictions (mean) and human *Delivery* scores.

Metric	Accuracy	Fluency	Prosody	Total
Spearman	0.654	0.608	0.627	0.632

4 Task Fulfilment Evaluation

Our approach for *Task Fulfilment* evaluation combines a high-performing ASR system with a LLM guided by customized prompts to assess the completeness and relevance of the answer. Figure 1 illustrates the evaluation pipeline.

- **Automatic Speech Recognition (ASR) Module:** This component transcribes the spoken responses into text. We compared the performance of *Whisper* [9], *WhisperX* [10], *Google Speech-to-Text* [11], and *Wav2Vec 2.0 for CTC* [12] on the *EpaDB* dataset to evaluate their accuracy in transcribing relevant non-native speech. *WhisperX* had the best performance, achieving a *Word Error Rate (WER)* of 8.13%.
- **Large Language Model (LLM) with Prompting:** This component evaluates the semantic content of our transcriptions using *LLaMA 3 8B* [13], a model fine-tuned for instruction. For each test part (Sect. 2.1), we design customized prompt sets to assess the degree of *Task Fulfilment*. We selected this model because it can be run locally, which ensures data privacy and facilitates unrestricted experimentation during development. While larger commercial models may achieve better performance, our modular pipeline design allows for their future integration.

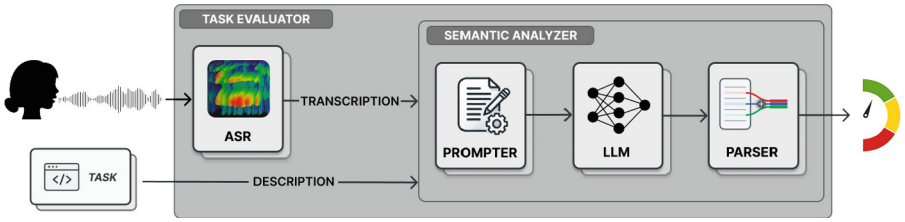


Fig. 1. Evaluation pipeline for assessing *Task Fulfilment*.

4.1 Dataset and Annotations

Our evaluation employs a dataset derived from the *UST Dataset* (Sect. 2.1), which includes responses annotated with human-assigned *Task Fulfilment* scores. While the original dataset provided one global score per category at the student

level, our new dataset is structured at response level, with each row corresponding to an individual answer given to a specific task. This restructuring allows for a more granular analysis of response variability.

Data Augmentation. Building upon the original *UST Dataset*'s 13 tasks, we manually developed 21 additional ones, resulting in a broader set of 34 tasks across three categories. To support this expansion, we included both **manually curated responses** and **LLM-generated answers** from models such as *GPT-4* [14] and *Claude 3.5 Sonnet* [15]. Approximately 45% of the responses were sourced from the original dataset, 17% were manually authored by our team, and 38% were generated using LLMs. All curated responses were manually validated by annotators to ensure consistency and task relevance, with disagreements resolved collaboratively. The selected responses reflect the linguistic range and answer styles found in the original dataset and are used to study the Semantic Analyzer module in isolation. Since our pipeline is modular and the ASR component was previously validated, analyzing textual inputs directly provides meaningful insight into the scoring process while enabling a broader evaluation across tasks.

Annotation and Scoring Criteria. For each response in the new dataset, we assigned one of three ordered labels based on its ability to address the prompt: a response is considered **Not Addressed** if it fails to provide adequate or relevant content, **Partially Addressed** if it covers some aspects of the prompt but remains incomplete or contains significant errors that hinder complete fulfilment, and **Fully Addressed** if it effectively and comprehensively meets the task requirements, conveying the intended meaning even with minor language errors. This three-level scoring system captures subtle differences in response quality and enables direct comparisons between automated evaluations and human judgments. Across the full set of annotated responses, 42% were labeled as *Fully Addressed*, 25% as *Partially Addressed*, and 33% as *Not Addressed*.

4.2 Evaluation Strategy and Results

Our evaluation of *Task Fulfilment* is organized by test part, recognizing that each has its own specific requirements. For each, we developed a set of custom prompt templates that incorporate the task details and the corresponding student response. This framework provides the necessary context to assess whether the answer is complete and relevant.

To build the prompts, we experimented with **few-shot** learning [16] and **chain of thought** prompting [17]. We refined all the prompts with clear instructions and consistent output formatting (JSON) so that the model can reliably distinguish among the defined categories. Furthermore, to ensure that the evaluation focuses on the intended meaning rather than penalizing non-critical language mistakes, we explicitly encouraged tolerance toward common *English as*

a Second Language (ESL) variations such as minor grammatical, lexical, and syntactic errors.

Table 2 summarizes the best results obtained for each test part, using Cohen's weighted Kappa to quantify agreement with human ratings. These results highlight distinct patterns across the different parts of the test. For Part 1, prompts with concrete examples proved especially effective in reducing ambiguity when handling brief or underspecified student answers, which are common in short-response settings. Part 2, which relies on interpreting visual input, posed a challenge due to the model's lack of native vision. To mitigate this, we included detailed textual descriptions of the images in the prompt; in this setting, the best results were achieved by combining few-shot prompting with chain of thought reasoning. Nonetheless, the model still struggled with partially relevant answers, suggesting limits in nuance handling. In Part 3, chain of thought prompting helped ensure that all expected elements were covered in longer responses, supporting more consistent judgments in complex settings.

To further assess the reliability of our best-performing prompts for Part 1 tasks, we ran each prompt 30 times across the dataset to account for the stochastic nature of LLM outputs. Few-shot prompting showed high stability (average pairwise Cohen's weighted Kappa ≈ 0.96), while a combined approach integrating chain of thought slightly improved agreement but increased variability. These results suggest that incorporating concrete examples helps stabilize the model's responses.

Table 2. Agreement scores by test part and prompting strategy.

Part	Prompting Strategy	Kappa
1	Few-shot	0.83
2	Few-shot + Chain-of-thought	0.82
3	Chain-of-thought	0.86

5 Conclusions and Future Work

This study examined automated methods for assessing spoken responses in a large-scale educational setting in Uruguay, focusing on *Delivery* and *Task Fulfilment*. Our results show that state-of-the-art speech processing models and LLMs can match human judgments with high agreement. *MultiPA* exhibited significant correlations with *Delivery* scores, while prompting strategies for *Task Fulfilment* achieved strong alignment, with weighted Kappa scores above 0.81.

Dataset expansion efforts are underway to increase the amount of available speech data and introduce other per-audio annotations, allowing for a more granular analysis of spoken responses across both dimensions. Additionally, improving data collection strategies, such as minimizing background noise and refining

preprocessing techniques to better filter non-English content, could enhance system robustness and ensure higher-quality inputs for automated assessment.

Further improvements can be made in the classification of *Task Fulfilment* responses, particularly in distinguishing borderline cases where *Partially Addressed* answers pose challenges. Future work may explore better LLMs and improved prompt engineering techniques to enhance classification accuracy.

Future research on *Delivery* prediction would benefit from per-audio annotations, enabling a more detailed evaluation of *MultiPA* and alternative speech assessment models. Additionally, if sufficient annotated data becomes available, fine-tuning *MultiPA* could be explored as a promising approach. Adapting its final layers while preserving core representations may help mitigate cross-linguistic mismatches and improve alignment with human assessments.

References

1. Levy, M.: Computer-Assisted Language Learning: Context and Conceptualization. Clarendon Press, Oxford, England (1997)
2. Peterson, M.: Digital simulation games in CALL: a research review. Comput. Assist. Lang. Learn. **36**, 943–967 (2023). https://doi.org/10.1080/09588221.2021.1954954
3. Zechner, K., Evanini, K. (eds.): Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech, 1st edn. Routledge (2019). https://doi.org/10.4324/9781315165103
4. Ceibal en Inglés. https://ceibal.edu.uy/plataformas-y-programas/ceibal-en-ingles/
5. Khabbazbashi, N., Nakatsuhara, F., Inoue, C., Kaplan, G., Green, A.: The design and validation of an online speaking test for young learners in Uruguay: challenges and innovations. Int. J. TESOL Stud. (2022). https://doi.org/10.46451/ijts.2022.01.10
6. Vidal, J., Ferrer, L., Brambilla, L.: Epadb: a database for development of pronunciation assessment systems. In: Proceedings of Interspeech 2019, pp. 589–593. ISCA (2019)
7. Kyriakopoulos, K.: Deep learning for automatic assessment and feedback of spoken English. Ph. D. thesis, University of Cambridge (2021)
8. Chen, Y.W., Yu, Z., Hirschberg, J.: Multipa: a multi-task speech pronunciation assessment model for open response scenarios. In: Proceedings of Interspeech 2024, pp. 297–301. ISCA (2024)
9. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. arXiv:2212.04356 (2022)
10. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: time-accurate speech transcription of long-form audio. In: Proceedings of the 24th Interspeech Conference, pp. 4489–4493. International Speech Communication Association (ISCA) (2023)
11. Google cloud speech-to-text. https://cloud.google.com/speech-to-text
12. Baevski, A., Zhou, H., 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)
13. Grattafiori, A., Dubey, A., Jauhri, A., et al.: The Llama 3 herd of models. arXiv:2407.21783 (2024)

14. OpenAI, Achiam, J., Adler, S., et al.: GPT-4 technical report. arXiv:2303.08774 (2023)
15. Anthropic: Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet
16. Brown, T.B., Mann, B., Ryder, N., et al.: Language models are few-shot learners. arXiv:2005.14165 (2020)
17. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, pp. 24824–24837. Curran Associates, Inc., Red Hook, NY, USA (2022)



Efficient Multi-task Inferencing: Model Merging with Gromov-Wasserstein Feature Alignment

Luyang Fang[1,3], Ehsan Latif[1,2], Haoran Lu[3], Yifan Zhou[1,4], Ping Ma[3], and Xiaoming Zhai[1,2(✉)]

[1] AI4STEM Education Center, Athens, GA, USA
xiaoming.zhai@uga.edu
[2] Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA
[3] Department of Statistics, University of Georgia, Athens, GA, USA
[4] School of Computing, University of Georgia, Athens, GA, USA

Abstract. Automatic scoring of student responses enhances efficiency in education, but deploying a separate neural network for each task increases storage demands, maintenance efforts, and redundant computations. To address these challenges, this paper introduces the Gromov-Wasserstein Scoring Model Merging (GW-SMM) method, which merges models based on feature distribution similarities measured via the Gromov-Wasserstein distance. Our approach begins by extracting features from student responses using individual models, capturing both item-specific context and unique learned representations. The Gromov-Wasserstein distance then quantifies the similarity between these feature distributions, identifying the most compatible models for merging. Models exhibiting the smallest pairwise distances, typically in pairs or trios, are merged by combining only the shared layers preceding the classification head. This strategy results in a unified feature extractor while preserving separate classification heads for item-specific scoring. We validated our approach against human expert knowledge and a GPT-o1-based merging method. GW-SMM consistently outperformed both, achieving higher micro F1 score, macro F1 score, exact match accuracy, and per-label accuracy. The improvements in micro F1 and per-label accuracy were statistically significant compared to GPT-o1-based merging ($p = 0.04$, $p = 0.01$). Additionally, GW-SMM reduced storage requirements by two-thirds without compromising much accuracy, demonstrating its computational efficiency alongside reliable scoring performance.

Keywords: Multi-task learning · Efficient inference · Automatic Scoring · Deep Learning · Optimal Transport

L. Fang and E. Latif—Co-first authors.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 192–200, 2025.
https://doi.org/10.1007/978-3-031-99267-4_24

1 Introduction

Automated scoring systems have become indispensable in modern educational assessment, enabling efficient evaluation of students' open-ended responses, particularly in science and STEM domains [15,17]. As curriculum frameworks such as the Next Generation Science Standards (NGSS) promote complex performance tasks to assess multidimensional understanding [9], the demand for reliable and scalable automated scoring continues to rise. Deep learning models like BERT [3], offer robust performance in such tasks. However, deploying a separate fine-tuned model for each assessment item is often impractical due to high memory and inference costs—especially in resource-constrained environments such as browser-based student assessments or serverless educational platforms [29].

In large-scale assessments, models are frequently fine-tuned separately per task to align with domain-specific rubrics and expert scoring criteria [22]. Over time, educational organizations may accumulate numerous task-specific models. Unlike multitask learning setups where tasks are trained jointly from scratch, re-training or co-training becomes infeasible due to data silos, annotation inconsistencies, and vendor diversity [14]. While parameter-efficient tuning methods like LoRA [11], adapters [4], or knowledge distillation [5,8] offer alternatives, they require retraining access and architectural constraints that are not always practical in real-world deployments.

The problem we address is thus real and pressing: *How can we consolidate multiple pre-trained scoring models, each trained on different tasks, into a smaller set of merged models without causing significant performance degradation?* This scenario frequently arises in automated essay scoring (AES) and short-answer grading (ASAG), particularly in cross-prompt trait scoring settings [12,28]. Prior works have explored joint training across prompts, transfer learning across similar tasks [16], and meta-learning for prompt generalization [7], but the question of post hoc model merging remains largely underexplored.

To bridge this gap, we introduce *Gromov-Wasserstein Scoring Model Merging (GW-SMM)*, a new algorithm that merges pre-trained models by aligning their learned response representations. In contrast to prior methods that rely on parameter similarity or prompt metadata, GW-SMM uses the Gromov-Wasserstein (GW) distance, a structure preserving optimal transport metric [19,24], to measure alignment between model-specific student response features. This strategy ensures that only models with structurally compatible internal representations are merged. It minimizes conflicts and maximizes knowledge transfer.

The proposed method supports resource-efficient AI for education and other cost-sensitive domains, contributing to advancements in scalable, efficient model deployment. Our key contributions are summarized below:

– We propose a novel model merging approach using GW distance to align deep semantic representations of student responses, enabling scalable deployment of scoring models without retraining.

– We empirically validate GW-SMM on NGSS-aligned science assessments across nine tasks, demonstrating significant storage reduction with minimal performance trade-offs, outperforming both human knowledge-based and GPT-derived merging strategies.
– We also open-source the code on Github[1] repository for reproducibility.

2 Gromov-Wasserstein Scoring Model Merging

We propose Gromov-Wasserstein Scoring Model Merging (GW-SMM) to merge multiple task-specific scoring models into a reduced set while preserving performance. Given T fine-tuned models trained on distinct NGSS-aligned science assessments, GW-SMM merges models by aligning their learned semantic features of student responses. The workflow is shown in Fig. 1.

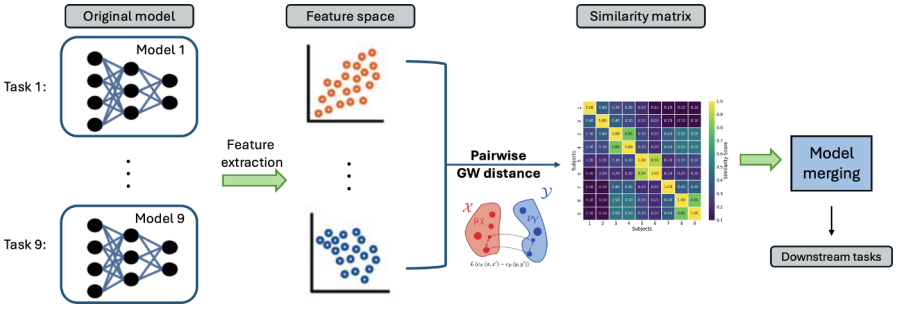


Fig. 1. Overview of proposed GW-SMM, where models for tasks are merged based on the inter-feature GW distance.

Task-Specific Representation Extraction. For each fine-tuned automated scoring model, we extract high-dimensional semantic features of student responses by passing task-specific data through the corresponding model and collecting the activations from the final hidden layer (pre-classification). In transformer architectures (e.g., BERT), this layer captures task-specific semantic patterns critical for scoring decisions, as shown in prior work on model similarity analysis [3,20].

Structural Alignment via Gromov-Wasserstein Distance. To identify which models should be merged together, we assess similarity based on their extracted features, which reflect both student response information and underlying model structures. We compute pairwise distances between their feature spaces using the GW distance. The GW distance captures the geometric alignment of features, allowing comparison even when models encode responses in heterogeneous metric spaces, such as those with different dimensionalities or context-dependent semantics.

[1] https://github.com/AI4STEM-Education-Center/MoE.git.

Formally, let $\mathbf{X} \in \mathbb{R}^{n_i \times d_i}$ and $\mathbf{Y} \in \mathbb{R}^{n_j \times d_j}$ denote the extracted feature for model i and j, where n_i, n_j and d_i, d_j may differ across models. Note that we work on a special case that each fine-tuned model has the same model architecture, i.e., $d_i = d_j$, but the framework is generally feasible for settings with different d_i and d_j. Let $C_i \in \mathbb{R}^{n_i \times n_i}$ and $C_j \in \mathbb{R}^{n_j \times n_j}$ denote intra-model cost matrices where $C_i[k,l] = distance_i(\mathbf{x}_k, \mathbf{x}_l)$ and $C_j[m,n] = distance_j(\mathbf{y}_m, \mathbf{y}_n)$. To align the feature structures between two models, we introduce a coupling matrix $\pi \in \mathbb{R}_+^{n_i \times n_j}$, where $\pi_{k,m}$ indicates the matching strength between the k-th point in $\mathbf{X}$ and the m-th point in $\mathbf{Y}$. The GW alignment is computed through the coupling matrix $\pi \in \mathbb{R}_+^{n_i \times n_j}$ solving:

$$GW(C_i, C_j)^2 = \min_{\pi \in \Pi(\mathbf{p},\mathbf{q})} \sum_{k,l=1}^{n_i} \sum_{m,n=1}^{n_j} \pi_{k,m}\pi_{l,n} L(C_i[k,l], C_j[m,n]), \tag{1}$$

where L is a distance measure with common choices and $L(a,b) = |a-b|^p$, $p \geq 1$, and $\Pi(\mathbf{p},\mathbf{q}) = \left\{\pi \mid \pi_{k,m} \geq 0; \forall k, \sum_{m=1}^{n_j} \pi_{k,m} = p_k; \forall m, \sum_{k=1}^{n_i} \pi_{k,m} = q_m\right\}$ enforces marginal constraints with $\mathbf{p} \in \mathbb{R}^{n_i}, \mathbf{q} \in \mathbb{R}^{n_j}$ as weights for points in $\mathbf{X}$ and $\mathbf{Y}$ (typically uniform: $p_k = 1/n_i, q_m = 1/n_j$).

Direct optimization of GW is challenging due to its non-convex and non-smooth properties. Thus, we use entropy regularization [23], which renders the problem convex while maintaining geometric fidelity between feature spaces, with the optimal coupling computed via Sinkhorn iterations [2,25]. For theoretical guarantees, computational considerations, and extensions of the discrete GW formalism, we direct readers to [21,24].

Determination of the Merging Plan. Given the distance matrix and the target number T' of final merged models, we determine the optimal merging plan by maximizing the total similarity, $L(\mathcal{M}) = \sum_{(i,j)\in\mathcal{M}} S_{i,j}$, where $\mathcal{M}$ denotes the set of merged pairs or groups, and $S_{i,j}$ is the normalized similarity between models i and j, defined as $S_{i,j} = 1 - \frac{d_{i,j} - d_{\min}}{d_{\max} - d_{\min}}$, where $d_{i,j}$ is the GW distance between models i and j, and $d_{\min}$, $d_{\max}$ are the minimum and maximum distances. We search for the optimal merging plan $\mathcal{M}^*$ that minimizes $L(\mathcal{M})$ over all possible merge plans that will result in T' final merged models.

Model Merging. With the designed merging plan, we adopt `TIES-MERGING` [28] to merge task-specific models into a unified model while preserving performance. Suppose we have T models that will be merged together. For model $t \in \{1, \cdots, T\}$, we compute the parameter updates $\tau_t = \theta_t - \theta_0$, where θ_t denotes the parameters for the task-specific model t while θ_0 denotes the parameters for the base model. With an alignment and pruning process to remove redundant parameters and conflicting parameters across different models via magnitude-based sparsification, we have the updated updates γ_t, and the final merged parameters are $\theta_{\text{merged}} = \theta_0 + \sum_{t=1}^{T} \lambda_t \gamma_t$, where the coefficients λ_t are determined by the alignment and pruning process. By combining representation alignment with targeted pruning, `TIES-MERGING` effectively leverages shared knowledge across tasks while maintaining task-specific distinctions.

3 Dataset Details

This study utilizes expert-scored responses from approximately 1,200 middle school students across the U.S., completing nine NGSS-aligned, open-ended science assessment tasks from the PASTA project [9,22]. These multi-label tasks, adapted from the Next Generation Science Assessment (NGSA) initiative, target physical science concepts under "Matter and Its Characteristics" domain and require students to integrate disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and science and engineering practices (SEPs) [1]. Responses were anonymized and collected in diverse educational contexts, reflecting variability in instructional methods, digital access, and linguistic backgrounds. For example, in one task, students analyzed density and flammability data to identify unknown gases, applying SEP, CCC, and DCI reasoning. All responses were scored using a structured multi-dimensional rubric co-designed with educators, enabling nuanced evaluation of students' scientific thinking [10]. Detailed dataset statistics and scoring rubrics are available on the project repository[1].

4 Experimentation and Results

For each task, we have a fine-tuned BERT-base model [3] for multi-label classification of student responses, tokenized via WordPiece [27]. Each response was mapped to a fixed-length sequence with [CLS]/[SEP] tokens and encoded by BERT. Training used AdamW [18] (learning rate: $2e^{-5}$, batch size: 32) with binary cross-entropy loss per label and dropout regularization [26]. Early stopping was applied based on validation loss. The training process was conducted for varying numbers of epochs (ranging from 10 to 20) on an NVIDIA GPU. For merging, we removed the classification head and extracted [CLS] token embeddings (dimension $d = 768$) from the final layer, forming a matrix (n_i, d) for each task, where n_i is the total number of student responses. Pairwise Euclidean distances between features were computed to inform model merging decisions.

We evaluate each method using four multi-label metrics: Micro F1 (overall performance weighted by label frequency), Macro F1 (equal weight across labels), Exact Match (all labels must match exactly), and Per-label Accuracy (accuracy per label across instances). Together, they capture both overall and label-specific performance. Performance is averaged across nine tasks, using merged or original models as applicable. We compare GW-SMM with two baselines: **(1) Human Knowledge:** Merging plans based on task similarities identified by domain experts through content and rubric analysis - reflecting common practices in educational assessment; **(2) GPT-o1:** Merging plans are generated by prompting GPT-o1 to compare task descriptions, inspired by LLM-based semantic similarity evaluation methods. As shown in Gatto et al. [6], such methods achieve strong performance in similarity tasks, making them reliable benchmarks.

We note that GW-SMM is a data-driven approach that leverages all student responses to infer task relationships, whereas Human Knowledge and GPT-o1 rely solely on task descriptions. This distinction stems from practical and computational constraints: human experts cannot feasibly analyze large volumes of

student responses manually, and GPT-o1 faces similar limitations due to data privacy concerns and the high cost of processing large-scale data [13].

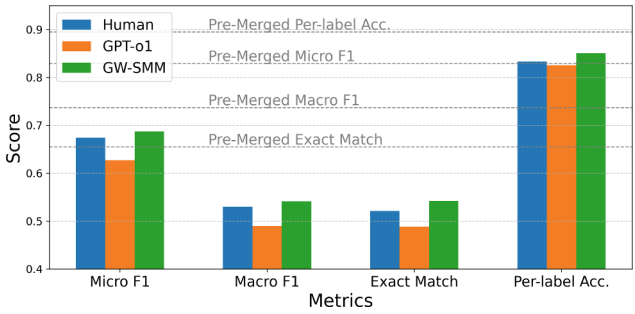


Fig. 2. Performance results before and after merging using three methods (GW-SMM (Ours), human knowledge, and GPT-o1).

Table 1. Statistical comparison of methods across metrics. Each value shows the t-statistic and p-value for pairwise comparisons among GW-SMM, Human Knowledge, and GPT-o1. Statistically significant results ($p < 0.05$) are bolded.

Metrics	GW-SMM vs Human		GW-SMM vs GPT-o1		Human vs GPT-o1	
	T-Stat	P-Value	T-Stat	P-Value	T-Stat	P-Value
Micro F1 Score	0.869983	0.391454	**2.106563**	**0.043925**	1.190602	0.243465
Macro F1 Score	-0.187336	0.852702	1.273012	0.213125	1.604841	0.119365
Exact Match Accuracy	1.363881	0.183098	1.173100	0.250302	0.062403	0.950670
Per-label Accuracy	1.577693	0.125483	**2.637638**	**0.013281**	0.769699	0.447703

In this study, we merged original models into $T' = 3$ merged models, reducing storage costs by 3x. As shown in Fig. 2, GW-SMM outperforms other methods, effectively aggregating models while retaining accuracy. Despite degradation versus the Pre-Merge baseline, GW-SMM achieves the optimal balance of efficiency and accuracy. Merging strategies and full results across metrics and tasks are available in our repository[5].

We also conducted paired t-tests to compare the effectiveness of each method (Table 1). The results reveal that GW-SMM and GPT-o1 exhibit statistically significant differences on two metrics: Micro F1 Score ($t = 2.107$, $p = 0.044$) and Per-label Accuracy ($t = 2.638$, $p = 0.013$). These results indicate that GW-SMM outperforms GPT-o1 in these aspects. In contrast, no statistically significant differences ($p > 0.05$) were observed between GW-SMM and Human Knowledge or between Human Knowledge and GPT-o1 across any metric. For instance,

comparisons between GW-SMM and Human Knowledge yielded p-values ranging from 0.126 to 0.853, while Human Knowledge versus GPT-o1 resulted in p-values between 0.120 and 0.951. These findings suggest that GW-SMM effectively leverages model merging to enable robust and efficient scoring, performing comparably to, or even exceeding, Human Knowledge in some cases.

5 Conclusion

We propose GW-SMM, a novel model merging method for the scalable deployment of commonly featured automatic scoring models. The method uses the Gromov-Wasserstein distance to align task-specific feature spaces by measuring structural divergence in student response features. GW-SMM identifies compatible models for merging while minimizing conflicts. Evaluated on nine NGSS-aligned science tasks, GW-SMM outperformed both human expert-guided and GPT-o1-based merging strategies, achieving statistically significant improvements in micro F1 score (0.6872 vs. 0.6271, $p = 0.04$) and per-label accuracy (0.8507 vs. 0.8255, $p = 0.01$). The method reduced storage costs by up to 3× while maintaining competitive performance relative to standalone models. Its data-driven approach, which leverages response patterns rather than task descriptions, ensures adaptability to diverse student responses. Future work will focus on advanced fusion techniques to further narrow the performance gap with pre-merged models. GW-SMM advances scalable, efficient deployment of AI in education, balancing accuracy and resource constraints.

Acknowledgments. This work was partially supported by the Institute of Education Sciences (IES) [R305C240010], the U.S. National Science Foundation (NSF) [DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809], and the National Institutes of Health [NIH R01GM152814]. Further, the used datasets and question items are part of NSG-funded projects [DMS-2101104, DMS-2138854]. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NIH, or IES.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Council, N.R., et al.: Next generation science standards: for states, by states (2013)
2. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems, vol. **26** (2013)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: arXiv preprint arXiv:1810.04805 (2018)
4. Ding, N., et al.: Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat. Mach. Intell. **5**(3), 220–235 (2023)
5. Fang, L., Chen, Y., Zhong, W., Ma, P.: Bayesian knowledge distillation: a Bayesian perspective of distillation with uncertainty quantification. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)

6. Gatto, J., Sharif, O., Seegmiller, P., Bohlman, P., Preum, S.M.: Text encoders lack knowledge: Leveraging generative LLMs for domain-specific semantic textual similarity. arXiv preprint arXiv:2309.06541 (2023)
7. Geden, M., Emerson, A., Rowe, J., Azevedo, R., Lester, J.: Predictive student modeling in educational games with multi-task learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 654–661 (2020)
8. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. Int. J. Comput. Vis. **129**(6), 1789–1819 (2021)
9. Harris, C.J., Krajcik, J.S., Pellegrino, J.W.: Creating and using instructionally supportive assessments in NGSS classrooms. NSTA Press (2024)
10. He, P., Shin, N., Zhai, X., Krajcik, J.: Guiding teacher use of artificial intelligence-based knowledge-in-use assessment to improve instructional decisions: a conceptual framework. In: Zhai, X., Krajcik, J. (eds.) Uses of Artificial Intelligence in STEM Education. Oxford University Press (2024)
11. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
12. Katuka, G.A., Gain, A., Yu, Y.Y.: Investigating automatic scoring and feedback using large language models. arXiv preprint arXiv:2405.00602 (2024)
13. Koedinger, K.R., D'Mello, S., McLaughlin, E.A., Pardos, Z.A., Rosé, C.P.: Data mining and education. Wiley Interdisc. Rev. Cogn. Sci. **6**(4), 333–353 (2015)
14. Koprinska, I., et al.: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings. Part II, Springer Nature (2023). https://doi.org/10.1007/978-3-031-74640-6
15. Latif, E., Zhai, X.: Fine-tuning ChatGPT for automatic scoring. Comput. Educ. Artif. Intell. **6**, 100210 (2024)
16. Latif, E., Zhai, X.: Efficient multi-task inferencing with a shared backbone and lightweight task-specific adapters for automatic scoring. In: AAAI Workshop on iRAISE: Innovation and Responsibility in AI-Supported Education, pp. 1–12. PMLR (2025)
17. Lee, G.G., Latif, E., Wu, X., Liu, N., Zhai, X.: Applying large language models and chain-of-thought for automatic scoring. Comput. Educ. Artif. Intell. **6**, 100213 (2024)
18. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Mémoli, F.: The Gromov-Wasserstein distance: a brief overview. Axioms **3**(3), 335–341 (2014)
20. Merchant, A., Rahimtoroghi, E., Pavlick, E., Tenney, I.: What happens to BERT embeddings during fine-tuning? arXiv preprint arXiv:2004.14448 (2020)
21. Panaretos, V.M., Zemel, Y.: Statistical aspects of Wasserstein distances. Ann. Rev. Stat. Appl. **6**(1), 405–431 (2019)
22. Pasta, P.T.: Supporting instructional decision making: potential of an automatically scored three-dimensional assessment system. https://ai4stem.org/pasta/ (2023)
23. Peyré, G., Cuturi, M., Solomon, J.: Gromov-Wasserstein averaging of kernel and distance matrices. In: International Conference on Machine Learning, pp. 2664–2672. PMLR (2016)
24. Peyré, G., Cuturi, M., et al.: Computational optimal transport: with applications to data science. Found. Trends® Mach. Learn. **11**(5-6), 355–607 (2019)
25. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pac. J. Math. **21**(2), 343–348 (1967)

26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learning Rres. **15**(1), 1929–1958 (2014)
27. Wu, Y.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
28. Yadav, P., Team: Ties-merging: a pruning-based approach to alleviate conflicts in model merging. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 4567–4578 (2023)
29. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. J. Res. Sci. Teach. **59**(10), 1765–1794 (2022)

Tutor or Mentor. A Comparative Usability Study of AI Roles in Higher Education

Luca Marconi[1(✉)] and Federico Cabitza[1,2]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
luca.marconi@unimib.it
[2] IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy

Abstract. Artificial Intelligence (AI) is transforming education, with Large Language Models (LLMs) increasingly supporting student learning through academic assistance, research facilitation, and engagement. However, the impact of different AI roles in educational contexts remains underexplored, particularly in terms of usability and effectiveness. Building on the AI role classification proposed by Mollick and Mollick (2023)—which differentiates between AI as a Tutor and AI as a Mentor—this study systematically evaluates their usability in student-AI interaction. We conducted a controlled experiment with 54 undergraduate students, comparing these AI roles across three key usability constructs: effectiveness, efficiency, and satisfaction. Results indicate that Tutor AI significantly enhances task effectiveness and demonstrates greater efficiency compared to Mentor AI, while satisfaction levels remain comparable across conditions. Additionally, a negative correlation between teamwork predisposition and user satisfaction emerges. These findings underscore the importance of tailoring AI pedagogical roles to diverse learning needs, with implications for the design of adaptive AI systems that balance structured guidance and self-regulated learning.

Keywords: Artificial Intelligence in Education · Large Language Models · Human-AI Interaction · Usability · Interaction Quality

1 Introduction

Artificial Intelligence (AI) is playing an increasingly prominent role in education, providing new opportunities for academic support, adaptive learning, and student engagement [3,5]. Large Language Models (LLMs), in particular, have shown significant potential in assisting students with academic tasks such as research, writing, and study organization [10,17]. AI-driven educational systems promise to enhance learning efficiency, facilitate knowledge acquisition, reduce cognitive load, and promote engagement [5,13], thereby supporting self-regulated learning [15]. Moreover, AI-powered tools can enhance educational accessibility and equity by supporting learners with diverse needs and skill levels [18].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 201–208, 2025.
https://doi.org/10.1007/978-3-031-99267-4_25

Despite the growing adoption of AI in education, a critical gap remains in understanding how varying AI interaction paradigms impact student engagement, usability, and learning effectiveness [14]. Prior research has often treated AI as a generic aid, without distinguishing between different functional roles [11]. However, AI systems can support learners in a variety of ways, reflecting a broad spectrum of pedagogical functions. Among these, a distinction emerges between systems providing real-time instructional scaffolding [2], while others act as reflective advisors, offering post-task feedback [19]. These approaches parallel pedagogical theories of guided versus self-regulated learning, yet there is limited empirical evidence comparing their usability.

To address this gap, we focus on a specific distinction proposed by Mollick and Mollick (2023), who classify AI roles as either *Tutor*—delivering real-time, step-by-step support—or *Mentor*—providing feedback after task completion [16]. This study empirically compares these two AI roles in a controlled academic setting, assessing student perceptions across the three ISO 9241-11 usability dimensions: effectiveness, efficiency, and satisfaction [6].

Understanding these distinctions is significant as AI becomes increasingly integrated into academic settings. By clarifying these usability differences, our findings contribute to the development of pedagogically aligned, student-centered AI systems that optimize human-AI collaboration while preserving critical thinking and independent problem-solving.

2 Methods

This study investigated how AI roles—*Tutor* and *Mentor*—affect students' perceptions of human-AI collaboration in academic tasks. A quantitative approach was adopted to rigorously assess participants' experiences. Participants were recruited from the undergraduate course 'Human-System Interaction' in Artificial Intelligence at the University of Milano-Bicocca. Participation was voluntary and approved by the course instructors. Undergraduate students ($N = 54$) worked in collaborative groups of three to four members, each group randomly assigned to a single AI role to ensure unbiased allocation and balanced conditions. All participants had comparable academic backgrounds and AI exposure, as they were enrolled in the same program. Random group assignment further supported the balance of collaborative abilities. To ensure an unbiased allocation and balanced group distribution, all group members interacted with the same AI role. Demographic characteristics, including academic background and employment status, are summarized in Table 1.

Table 1. Demographic characteristics of participants

Characteristic	Details
Total Participants	54
AI Roles	*Tutor:* 28 (51.85%), *Mentor:* 26 (48.15%)
Age (Years)	Mean = 21.85, SD = 2.03, Range: 20–30
Academic Level	Upper Secondary: 83.3%
	Bachelor's Degree: 9.3%
Employment Status	Part-time/Internship: 11.1%

Each group was assigned a university-relevant task, such as preparing a project, organizing study plans, or developing research materials. Tasks were chosen independently by the groups within predefined academic categories. All tasks were university-related and of comparable complexity; task types were monitored across AI roles to minimize biases. The study design was informed by Mollick and Mollick's (2023) framework [16], which explores differentiated AI roles in educational contexts to enhance engagement and learning.

To simulate realistic academic usage scenarios, groups were allowed to interact with one or more LLMs, including ChatGPT, Bing, Claude2, Phind, Bard, or Poe. While this approach reflects real-world usage and enhances ecological validity, it also introduces potential variability in AI responses due to differences in model capabilities. Each group received an initial standardized prompt that clearly defined the AI's assigned role and outlined the interaction procedure. This prompt ensured a structured engagement aligned with the designated role. After this initial setup, participants interacted with the AI autonomously, managing the exchange while adhering to the role constraints specified in the prompt.

To promote consistency, all groups were instructed to align their interactions with the assigned AI role (Tutor or Mentor), document and archive complete interaction logs for each LLM used, and provide individual evaluations of the AI's contributions. No strict controls were imposed on prompt strategies or model selection, in order to preserve authentic user behavior and assess AI usability under realistic conditions.

The study followed a three-phase design. In the *pre-interaction phase*, students independently completed the *Teamwork Scale* [4], assessing collaborative predisposition, and the *Trust in Automation Scale (TAS)* [7], measuring trust in AI systems. During the *interaction phase*, each group engaged with a designated AI role (Tutor or Mentor), using at least two LLMs to complete an academic task. Participants evaluated each model's contribution independently. In the *post-interaction phase*, all students completed the *User Experience Testing Questionnaire* [12], measuring ISO 9241-11 usability constructs. Participants responded to the 12 items rated on a 5-point Likert scale. Each item addressed a specific dimension of interaction: REQ (requirements satisfaction), APPEAL (frustration, inverted), EASY (ease of use), FAST (perceived correction speed, inverted), QUICK (task execution speed), NOT-HARD (lack of difficulty,

inverted), EFFECT (task effectiveness), USEFUL (usefulness, inverted), COMPL (completion), UNDERSTAND (perceived understanding), SATISF (overall satisfaction), and REUSE (intention to reuse). Items with negative phrasing were reverse-scored so that higher values consistently reflect more positive experiences.

To validate the usability constructs, psychometric analyses were conducted using pairwise correlations and Cronbach's Alpha. Three composite variables were retained, each with acceptable internal consistency ($\alpha \geq 0.60$): *Effectiveness* (EFFECT and COMPL, $\alpha = 0.78$), *Efficiency* (QUICK and UNDERSTAND, $\alpha = 0.62$), and *Satisfaction* (SATISFACTION and REUSE, $\alpha = 0.72$). The aggregation method was validated through expert consensus, aligning with prior research on AI usability assessment [1].

Statistical analyses were conducted using Python to compare usability constructs between the Tutor and Mentor conditions. Due to the small sample size ($N = 54$) and potential violations of parametric assumptions, a Mann-Whitney U test was employed to ensure robust inference. A bootstrapping procedure (1,000 resamples) was applied to estimate 95% confidence intervals for usability scores. Additionally, Spearman's rank correlation was used to examine the association between teamwork predisposition and user satisfaction.

Ethical considerations were carefully addressed throughout the study to ensure the responsible handling of participant data and the integrity of the research process. All participants provided their informed consent and their responses were anonymized and securely stored, with measures implemented to protect confidentiality and prevent unauthorized access.

3 Results

This study examined differences in three key usability constructs—*Efficiency*, *Effectiveness*, and *Satisfaction*—between two AI roles: *Tutor* and *Mentor*. To ensure robustness, both parametric (independent samples t-test) and non-parametric (Mann-Whitney U test) analyses were conducted, with bootstrapped confidence intervals (CIs) enhancing result reliability.

Table 2 summarizes the comparative analysis of the *Mentor* and *Tutor* roles across the three usability dimensions. The table presents mean scores, bootstrapped confidence intervals (95%), and p-values from statistical tests.

Table 2. Comparison of Mentor vs Tutor results along usability dimensions.

Metric	EFFICIENCY	EFFECTIVENESS	SATISFACTION
Mentor Mean	3.49	3.35	4.14
Mentor CI 95%	(3.25, 3.71)	(3.08, 3.64)	(3.83, 4.38)
Tutor Mean	3.86	4.06	4.25
Tutor CI 95%	(3.57, 4.14)	(3.84, 4.29)	(4.02, 4.48)
MannWhitney p-value	0.034	0.001	0.685

The findings indicate that *Effectiveness* was significantly higher for the *Tutor* role ($p < 0.001$), while *Efficiency* showed a significant trend favoring *Tutor* ($p = 0.033$). No significant differences emerged for *Satisfaction*. Figure 1 illustrates that *Tutor* consistently outperforms *Mentor* in both *Efficiency* and *Effectiveness*, with particularly pronounced differences in *Effectiveness*, while *Satisfaction* scores remain consistently high across both roles.

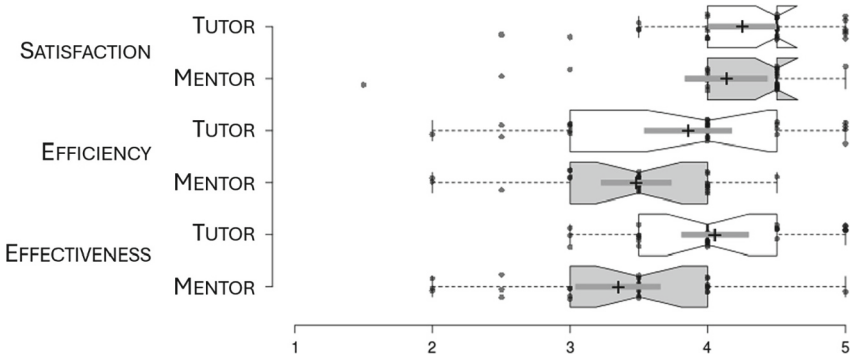


Fig. 1. Boxplots comparing usability constructs (Efficiency, Effectiveness, and Satisfaction) across AI roles (Mentor and Tutor). The crosses within the grey bars represent the means along with their 95% confidence intervals. The central bars within each boxplot denote the medians, while the notches indicate the 95% confidence intervals for the median.

The *Tutor* role exhibited higher *Efficiency* scores ($M = 3.86$, CI = [3.57, 4.14]) than the *Mentor* role ($M = 3.49$, CI = [3.25, 3.71]). Although the independent samples t-test did not reach statistical significance ($p = 0.061$), the *Mann-Whitney U* test suggested a significant trend favoring *Tutor* ($p = 0.033$). While these results suggest a potential efficiency advantage for *Tutor*, overlapping confidence intervals indicate that further research is needed to confirm its practical relevance.

A strong, statistically significant difference was observed in *Effectiveness*, with the *Tutor* role scoring significantly higher ($M = 4.06$, CI = [3.84, 4.29]) than the *Mentor* role ($M = 3.35$, CI = [3.08, 3.64]). The independent samples t-test ($p = 0.0005$) and *Mann-Whitney U* test ($p = 0.0012$) confirmed this effect. Satisfaction levels were high and consistent across both AI roles (*Mentor*: $M = 4.14$, CI = [3.83, 4.38]; *Tutor*: $M = 4.25$, CI = [4.02, 4.48]). No significant differences emerged ($p > 0.5$ for both tests), indicating that both roles provided a similarly positive user experience. Figure 2 presents the mean scores and corresponding 95% confidence intervals for all measured dimensions across the two AI roles.

A negative correlation emerged between students' teamwork predisposition and satisfaction with AI interactions ($r_s = -0.284$, $p = 0.045$). A bootstrapped test (10,000 iterations) confirmed the significance of this effect.

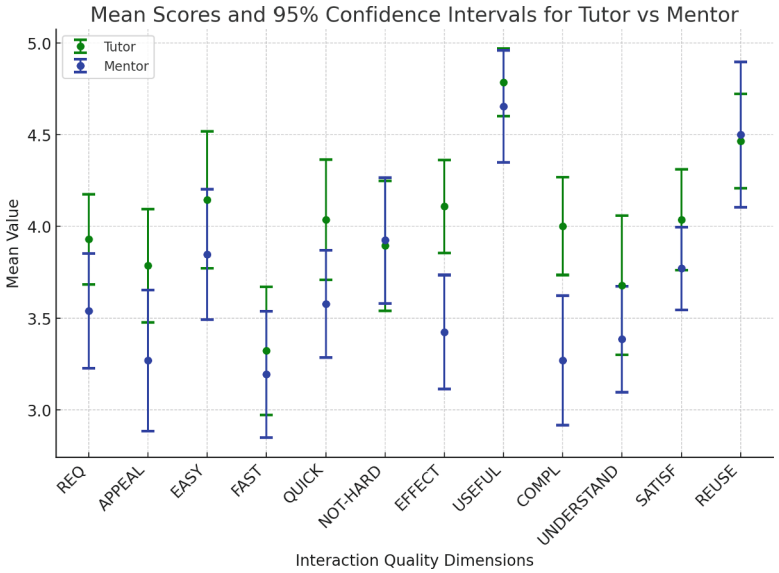


Fig. 2. Mean scores and 95% confidence intervals for interaction quality dimensions (Tutor vs Mentor).

4 Discussion and Conclusions

This study investigated how different AI pedagogical roles—Tutor and Mentor—influence students' perceptions of usability in academic collaboration, following the ISO 9241-11 framework (Effectiveness, Efficiency, Satisfaction).

Results revealed that students who interacted with the Tutor AI, which provides real-time guidance, reported significantly higher task effectiveness and slightly greater efficiency than those supported by the Mentor AI, who offered post-hoc feedback. These findings align with prior research emphasizing the benefits of structured scaffolding in educational contexts and real-time AI guidance in problem-solving scenarios [13]. Satisfaction levels were similarly high across both roles, suggesting that the *Mentor* AI may still play a valuable role in fostering metacognitive skills and reflective learning [8].

A noteworthy insight was the negative correlation between teamwork predisposition and AI satisfaction. Students with stronger inclinations toward human collaboration appeared less satisfied with AI interactions, possibly due to unmet expectations for social or dynamic engagement. This suggests a need for AI systems that integrate better with collaborative workflows, supporting—not replacing—peer interaction.

This study presents several limitations. First, it captures only short-term usability perceptions and task outcomes within a single session, leaving open whether the observed advantage of the Tutor AI supports long-term learning retention or skill transfer. While real-time scaffolding may enhance immediate

performance, prior research suggests it can hinder deeper conceptual engagement if it fosters overreliance on external support [9]. Second, the use of multiple LLMs without standardized control over model selection and prompt engineering may have introduced variability in AI responses across groups. Third, despite efforts to ensure comparable task complexity, the self-selected and non-identical tasks could have resulted in unbalanced distributions across conditions, potentially compromising comparability. Task variability may have influenced performance independently of the AI's role. Moreover, the sample size ($N = 54$) and the exclusive focus on undergraduate students limit the generalizability of the findings. Participants were drawn solely from a single course at one institution, which limits the applicability of the results to more diverse educational contexts. Finally, although Tutor AI's effectiveness was statistically confirmed, the study does not clarify the underlying cognitive mechanisms, such as whether gains stem from reduced cognitive load, increased motivation, or other factors. These limitations warrant caution in generalizing the findings beyond the specific context examined. Future research should investigate the robustness of these results across educational levels, disciplines, and institutions, ideally through longitudinal, mixed-methods studies employing standardized AI systems and tasks.

Despite these limitations, the findings offer design implications for educational AI. The Tutor AI's success supports the integration of adaptive, real-time feedback in learning environments. Yet, the comparable satisfaction across roles highlights the potential for hybrid AI systems that dynamically shift between Tutor-like and Mentor-like modes, optimizing support based on learner needs. Additionally, improving AI compatibility with team-based learning may enhance acceptance among students with strong collaborative preferences.

Future work should employ longitudinal, mixed-methods research to explore lasting learning impacts, better delineate AI-task dynamics, and promote equitable, personalized, and socially responsive AI integration in academic settings.

References

1. Alabduljabbar, R.: User-centric AI: evaluating the usability of generative AI applications through user reviews on app stores. PeerJ Comput. Sci. **10**, e2421 (2024). https://doi.org/10.7717/peerj-cs.2421
2. Alsobeh, A., Woodward, B.: Ai as a partner in learning: a novel student-in-the-loop framework for enhanced student engagement and outcomes in higher education. In: Proceedings of the 24th Annual Conference on Information Technology Education, pp. 171–172. SIGITE '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3585059.3611405
3. Brender, J., El-Hamamsy, L., Mondada, F., Bumbacher, E.: Who's helping who? When students use ChatGPT to engage in practice lab sessions. In: Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science, vol. 14829, pp. 235–249. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64302-6_17
4. French, B.F., Gotch, C.M., Immekus, J.C., Beaver, J.L.: An investigation of the psychometric properties of a measure of teamwork among high school students. Psychol. Test Assess. Model. **58**(3), 455 (2016)

5. Heung, Y., Chiu, T.K.: How ChatGPT impacts student engagement from a systematic review and meta-analysis study. Comput. Educ. Artif. Intell. **8**, 100361 (2025). https://doi.org/10.1016/j.caeai.2025.100361
6. International Organization for Standardization: ISO 9241-11:2018 Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts (2018)
7. Jian, J., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. Int. J. Cogn. Ergon. **4**(1), 53–71 (2000). https://doi.org/10.1207/S15327566IJCE0401_04
8. Jin, S.H., Im, K., Yoo, M., Roll, I., Seo, K.: Supporting students' self-regulated learning in online learning using artificial intelligence applications. Int. J. Educ. Technol. High. Educ. **20**(1), 37 (2023). https://doi.org/10.1186/s41239-023-00406-5
9. Johanna Lönngren, T.A., Svanström, M.: Scaffolding strategies in a rubric-based intervention to promote engineering students' ability to address wicked problems. Eur. J. Eng. Educ. **44**(1–2), 196–221 (2019). https://doi.org/10.1080/03043797.2017.1404010
10. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023). https://doi.org/10.1016/j.lindif.2023.102274
11. Kim, J., Yu, S., Detrick, R., Li, N.: Exploring students' perspectives on generative AI-assisted academic writing. Educ. Inf. Technol. **30**, 1265–1300 (2025). https://doi.org/10.1007/s10639-024-12878-7
12. Leong, P.H., Goh, O.S., Yogan, J.K., Sam, Y.H., Fong, C.W.: The evaluation of user experience testing for retrieval-based model and deep learning conversational agent. Int. J. Adv. Comput. Sci. Appl. **12**(4) (2021)
13. Looi, C.K., Jia, F.: Personalization capabilities of current technology chatbots in a learning environment: an analysis of student-tutor bot interactions. Educ. Inf. Technol. (2025). https://doi.org/10.1007/s10639-025-13369-z
14. Memarian, B., Doleck, T.: A multidimensional taxonomy for learner-AI interaction. Educ. Inf. Technol. **29**, 18361–18378 (2024). https://doi.org/10.1007/s10639-024-12546-w
15. Molenaar, I., de Mooij, S., Azevedo, R., Bannert, M., Järvelä, S., Gašević, D.: Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. Comput. Hum. Behav. **139**, 107540 (2023). https://doi.org/10.1016/j.chb.2022.107540
16. Mollick, E., Mollick, L.: Assigning AI: seven approaches for students, with prompts (2023)
17. Schön, E.M., Neumann, M., Hofmann-Stölting, C., Baeza-Yates, R., Rauschenberger, M.: How are AI assistants changing higher education? Front. Comput. Sci. **5**, 1208550 (2023)
18. Viberg, O., Kizilcec, R.F., Wise, A.F., Jivet, I., Nixon, N.: Advancing equity and inclusion in educational practices with AI-powered educational decision support systems (AI-EDSS). Br. J. Educ. Technol. **55**, 1974–1981 (2024). https://doi.org/10.1111/bjet.13507
19. Xu, X., Wang, X., Zhang, Y., Zheng, R.: Applying ChatGPT to tackle the side effects of personal learning environments from learner and learning perspective: An interview of experts in higher education. PLoS ONE **19**(1), e0295646 (2024). https://doi.org/10.1371/journal.pone.0295646

Workshops

XAI-Ed: Half-Day Workshop on Pedagogy-Founded Explainable AI for Transparent, User-Centered AI in Education

Hasan Abu-Rasheed[1(✉)], Mutlu Cukurova[2], Hassan Khosravi[3], Jeroen Ooge[4], and Christian Weber[5]

[1] Goethe University, Frankfurt, Frankfurt am Main, Germany
rasheed@studiumdigitale.uni-frankfurt.de
[2] University College London UCL, London, UK
[3] The University of Queensland, Brisbane, Australia
[4] Utrecht University, Utrecht, Netherlands
[5] University of Siegen, Siegen, Germany

Abstract. As educational technology evolves, integrating Artificial Intelligence (AI) into Technology-Enhanced Learning (TEL), Educational technologies (EdTech), and AI in Education (AIED) offer significant potential for learning and educational stakeholders. However, the opacity of many AI systems complicates the interpretation of their decision-making processes and their predictions of, e.g., personalized learning recommendations. Explainable AI (XAI) addresses this challenge by providing insights into how AI algorithms arrive at their predictions, thereby enabling stakeholders to make more informed decisions. This workshop will explore the convergence of XAI with AIED, TEL, and EdTech in general, aiming to empower learners and educators through transparent and understandable insights into the inner workings of AI algorithms. Although the recent advances within the AIED community underscored the growing demand for explainable and transparent AI systems, the focus has mainly been on the technical side of XAI, leaving a strong need for addressing the pedagogical aspects of XAI in education. Those are not only limited to the pedagogical foundations of XAI but also include the actual learning value from implementing XAI, how this value aligns with learning theories, and how XAI's ability to achieve it can be evaluated. In educational contexts, where learner autonomy and agency are essential for a fair, learner-centered process, XAI is expected to provide meaningful explanations for algorithmic predictions. Successfully implementing XAI in EdTech and AIED, therefore, requires a robust understanding of its pedagogical foundations, technical realization, regulatory limitations, and the perspectives of the educational stakeholders, including learners, educators, developers, and institutions.

Keywords: Explainable AI in Education · Pedagogy · Educational stakeholders · Human oversight · User-centered XAI

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 211–217, 2025.
https://doi.org/10.1007/978-3-031-99267-4_26

1 Introduction

The safe and effective implementation of AI in education is an overarching goal of all AI systems developed and evaluated by the AIED community. The rapid development of AI, especially the growing international race for building and owning advanced and open-source Gen-AI models, has shown the importance of safeguarding and revising the implementation of AI in educational settings, due to the sensitivity of the educational domain, and the concrete requirements that it reveals from the long history of learning theories and pedagogical foundation of learning, which are mostly not in the core of the current AI and Gen-AI hype.

XAI in education is one of the main tools that focuses on the transparent, safe, and user-centered implementation of AI in an educational setting. Studies have highlighted the importance of explainability in fostering trust and acceptance of AI-driven prediction among learners [3]. In the context of education, where learners' autonomy and agency are keys to accomplishing a fair and learner-centered educational process, XAI becomes essential for ensuring that learners have meaningful insights into, e.g., why certain learning recommendations are provided to them.

Furthermore, the integration of XAI into AIED and EdTech aligns with the broader efforts to promote digital literacy and critical thinking skills among learners. By demystifying the decision-making processes of AI algorithms, XAI empowers learners and educators to critically evaluate and interrogate system prediction, e.g., learning recommendations. This aligns with the goals of promoting lifelong learning and fostering a culture of inquiry and self-directed learning, as advocated by UNESCO's Education 2030 Framework for Action [4]. Moreover, the ethical dimension of XAI in education cannot be overlooked [1, 2]. As AI technologies increasingly shape educational experiences, it is imperative to address concerns related to algorithmic bias, fairness, and accountability [5]. By incorporating XAI principles into the design and deployment of AI-driven learning systems, stakeholders can mitigate the risks of perpetuating inequities and ensure that educational opportunities (of making sound decisions on selecting or following a learning path) are accessible to all learners, regardless of their backgrounds or abilities. XAI-Ed becomes especially relevant to the AIED community since it focuses on ensuring a balance between the different disciplines involved in building and implementing AI algorithms in education, foremost are the didactics, technology, and legal ones. This workshop, therefore, emphasizes the interdisciplinary design of XAI-Ed, prioritizing the role of educational stakeholders and balancing the development of XAI-Ed on its main pillars:

- its pedagogical foundation, e.g., in learning theories and learning design.
- its technical realization, in terms of the best approaches to achieve explainability of black-box AI models
- its regulatory framework, i.e., within the European AI Act, Data Privacy Act, and the similar regulations that govern AI implementation on different risk levels.
- its evaluation and role assessment, from the perspective of the actual learning outcome and the relevant stakeholders, to ensure that XAI brings real value to learning.

2 Format and Activities

Our XAI-Ed workshop is designed to serve as a forum for interdisciplinary dialogue and collaboration, bringing together researchers, developers, educators, and practitioners to explore the transformative potential of XAI and the best practices for its integration in education from the different perspectives of their disciplines.

Four main types of activities will be utilized in the workshop:

- Presentations
- Open discussions
- Group work
- Hands-on exploration

The workshop is structured to encourage active participation through group work and discussions, as well as collaborative work among participants through hands-on activities in which they explore existing XAI tools and analyze them according to the theoretical concepts introduced at the start of the workshop by the workshop chairs. Planned for the hands-on exploration are two applications, one focusing on the control features of XAI, and the second focuses on teacher-oriented experimentation with XAI. High interactivity is expected and will be encouraged throughout the workshop to foster engagement and knowledge exchange among participants.

In order to set a solid foundation for the discussions, the workshop program will utilize a set of presentations, the first of which is conducted by the workshop chairs, and the rest are from the authors of the papers accepted in the workshop, who will be given the time to present their research and answer potential questions from the audience about it.

3 Workshop Content and Theme

The theme of the XAI-Ed workshop is the proper design and implementation of XAI for education. It originates from the challenges that XAI faces if it is perceived as a purely technical process that is meant to be fitted into education. We highlight that the design decisions, implementation steps, and evaluation of the effects of XAI in education cannot be carried out by one discipline without the other, but rather as an interdisciplinary process, in which technical and educational stakeholders codesign the explainability of AI in education, based on theoretical and practical foundations from both domains. This, of course, does not exclude the role of regulatory stakeholders, who are also involved in the design process through the legislation frameworks that are developed on national and international levels.

To achieve this goal, the workshop's content is designed to cover the core topics needed to dive deeper into understanding XAI from its technical and didactical perspectives. This will provide food for thought, which participants will enrich and use in the discussions and presentations within the workshop. The planned content of the workshop and its format are shown in Table 1.

Table 1 Workshop's content and time plan

Session	Session Topic	Approach
Integration of explainable AI into AIED and EdTech (20min)	Welcome, introductions, and overview of workshop objectives and agenda	Presentation
	The principles and importance of explainable AI in education: • Why XAI? • XAI in education • XAI perspectives and stakeholders • Explainability vs. Interpretability vs. Transparency • XAI and FATE principles	Presentation
	Connection of education stakeholders to FATE principles and XAI perspectives Influence of different educational contexts, e.g., higher education or vocational education and training, on the design of XAI	Discussion
Accepted paper presentations (1h 10min)	4 min presentation + 3 min Q/A time slot for each accepted paper	Presentation
Break (30min)		
Methods, tools, and pedagogy foundation of XAI (40min)	State-of-the-art approaches for designing and evaluating the explainability of AI systems, especially in education	Presentation
	Challenges and opportunities of integrating XAI into AIED	Discussion
	Introducing and testing tools for XAI in education	Group-work
Group discussion and moderated open discussion (40min)	How can/should XAI be designed in your context? By whom? And for whom?	Group Discussion
	Ethical dilemmas that face XAI implementation Requirements expected from technology and pedagogy to fulfill and support an ethical design of XAI in education	Group Discussion
	Future directions and research opportunities for the field of XAI in AIED Role and agency of different stakeholders towards XAI in education Legal requirements, potentials, and limitations	Group Discussion
Closing, reflection, and future perspectives (10min)	Recap of key insights and takeaways from the workshop Overview of the workshop open resources Closing remarks and acknowledgments	

4 Workshop Organization

4.1 Organizing Team

The workshop is designed and conducted by a team combining interdisciplinary expertise and focused on advancing the research on XAI in education as well as the community supporting it. Workshop chairs are:

Prof. Mutlu Cukurova is a professor of Learning and Artificial Intelligence IOE - Culture, Communication & Media. He investigates the potential of analytics and artificial intelligence solutions to support human learning. His primary research interest lies in enhancing individuals' ability to learn how to learn and solve complex problems collaboratively. On a broader scale, his work aims to address the critical socio-educational challenge of preparing people for a future alongside AI agents, where skills beyond the routine cognitive abilities currently emphasized by many education systems will be essential.

Dr. Hassan Khosravi is an Associate Professor in Data Science and Learning Analytics at The University of Queensland. As a computer scientist by training, he is passionate about the role of artificial intelligence in the future of education. In his research, he draws on theoretical insights driven by the learning sciences and exemplary techniques from the fields of human-centered AI and crowdsourcing to build technological solutions that enhance student learning and experience. His past research and publications have addressed a number of diverse topics, such as learning graphical models, statistical-relational learning, social network analysis, cybersecurity, and game theory.

Dr. Jeroen Ooge is an assistant professor at Utrecht University. His research is situated in human–computer interaction, which investigates how people interact with technologies. Dr. Ooge specializes in human-centered explainable artificial intelligence, which studies how outcomes of AI models can be explained to different audiences in different contexts, for example, teenagers in an educational context. He is interested in how transparency affects people's trust in AI models, how it supports people's decision-making, whether it improves people's understanding of AI models, etc., with a particular focus on studying how data visualization can help in this endeavor.

Dr. Christian Weber is a substitute professor and postdoctoral researcher at the University of Siegen, Germany. He focuses on advancing digitalization in industry, education, and medicine through AI-driven, semantically enhanced knowledge maps, particularly in mobile health systems. Dr Weber leads the intelligent infrastructure division within the interdisciplinary "digital practice" research group, overseeing projects that integrate computer science, medicine, education, and manufacturing, all while emphasizing the essential role of the human factor in digital solutions.

Hasan Abu-Rasheed is a research associate in Goethe University Frankfurt. His main field of research is technology-enhanced learning, utilizing different technologies such as knowledge graphs, explainable AI (XAI), recommender systems, and conversational agents (Chatbots) to develop explainable decision support systems for educators and learners.

5 Workshop Timeline

The plan to implement the workshop, including the CfP, spans three months be-fore the conference and the following month after the conference. This is to allow sufficient time for the authors to submit their contributions and manage the conference booking and registration on time.

Table 2 shows the tasks planned for the organization of the workshop. Although it begins with the notification of acceptance, the preparation of the workshop's content is an ongoing work that spans more than the organizational tasks in Table 2.

Table 2 Timeline for conducting the workshop before and after the conference

Task	Starts on	Ends on
Publishing workshop's website	21.03.2025	27.03.2025
Launch call for papers	01.04.2025	18.05.2025
Paper reviews	22.04.2025	31.05.2025
Notification of acceptance	31.05.2025	31.05.2025
Paper camera-ready version	01.06.2025	30.06.2025
Preparation for CEUR proceedings substitution	01.07.2025	17.07.2025
Workshop proceedings submission to CEUR	28.07.2025	31.08.2025

6 Plan for Publishing the Workshop's Proceedings

XAI-Ed workshop will announce a call for position and research papers, allowing the researchers in the field to share their research on the overarching topics of XAI development for education, XAI integration in the educational settings, and the educational stakeholder oversight over XAI. A detailed list of subtopics will be announced for the interested authors to ensure the relevance of their submissions to the workshop. Submitted papers will undergo a peer review process with at least two reviews per submission to select the best papers in terms of quality, rigor, originality, relevance, and discussion of ethical and societal implications.

To ensure the visibility and accessibility of the papers published in the workshop, the workshop chairs will compile workshop proceedings and submit them for publishing as CEUR Workshop Proceedings. The open-access nature of CEUR workshop proceedings serves as a means to disseminate the research done by the authors of accepted papers, thus supporting open science.

7 Expected Target Audience

The expected number of participants in the workshop is between 15 and 30 participants, but our formats allow for upscaling if needed. We expect participants will have diverse backgrounds, which will benefit the discussions. Although participants from the technical and pedagogical domains are expected to be the majority of the attendees, the workshop will also encourage participation from regulatory, educational-business, and educational management backgrounds.

8 Previous Workshop Edition

The XAI-Ed workshop is the second edition of the XAI in education workshop series that the authors are implementing. The first version carried the name "XAI4TEL: Explainable AI for Interpretable, Learner-Driven, Educational Technologies"[1] and was offered at the

[1] https://sites.google.com/view/xai4tel.

Nineteenth European Conference on Technology Enhanced Learning (ECTEL 2024), which took place between the 16th and 20th of September 2024. This workshop had to take place virtually at the last minute because of severe weather conditions at the physical venue. We therefore couldn't present all prepared materials as planned.

This second edition of the workshop is designed utilizing the experiences from the first edition and extended with additional content on two main topics: 1) the pedagogical foundation, integration, and evaluation of XAI in education, and 2) the regulatory perspectives of XAI in education, especially after the European AI Act was put into effect in August 2024.

References

1. Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López De Prado, M., Herrera-Viedma, E., Herrera, F.: Connecting the dots in trustworthy Artificial Intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf. Fus. **99**, 101896 (2023). https://doi.org/10.1016/j.inffus.2023.101896
2. Memarian, B., Doleck, T.: Fairnessss, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: a systematic review. Comput. Educ. Artif. Intell. **5**, 100152 (2023). https://doi.org/10.1016/j.caeai.2023.100152
3. Ooge, J., Kato, S., Verbert, K.: Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In: 27th International Conference on Intelligent User Interfaces (IUI 2022), pp. 93–105. ACM, Helsinki, Finland (2022). https://doi.org/10.1145/3490099.3511140
4. UNESCO: Incheon Declaration: Education 2030: Towards Inclusive and Equitable Quality Education and Lifelong Learning for All. UNESCO Digital Library (2015). https://unesdoc.unesco.org/ark:/48223/pf0000233137
5. Veale, M., Van Kleek, M., Binns, R.: Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In: Proceedings of the 2018 CHI Conf. on Human Factors in Computing Systems (CHI 2018), pp. 1–14. ACM, Montreal, Canada (2018). https://doi.org/10.1145/3173574.3174014



Workshop on Epistemics and Decision-Making in AI-Supported Education

Ehsan Latif[1,2], Ninghao Liu[3], Gautam Biswas[4], Yue Yin[5], and Xiaoming Zhai[1,2(✉)]

[1] AI4STEM Education Center, Athens, GA, USA
xiaoming.zhai@uga.edu
[2] Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA
[3] School of Computing, University of Georgia, Athens, GA, USA
[4] Department of Computer Science, Vanderbilt University, Nashville, TN, USA
[5] College of Education, University of Illinois Chicago, Chicago, IL, USA

Abstract. The Epistemics and Decision-Making in AI-Supported Education workshop explores the transformative role of generative AI—particularly large language and multimodal models—in reshaping educational practices. As these technologies evolve, they offer significant opportunities to enhance personalized learning, support higher-order thinking, and improve assessment design and evaluation. This workshop brings together educators, researchers, and AI practitioners to discuss how AI can create adaptive, inclusive learning environments while addressing crucial concerns around explainability, trust, and fairness. Topics include the use of large foundation models for content generation, AI-based personalization, creativity in learning, ethical considerations, and the evolving role of teachers in AI-integrated classrooms. Through keynote talks, peer-reviewed papers, and interactive sessions, participants will share research, explore real-world applications, and collaborate on future solutions. Selected papers will be published in the AIED proceedings. The workshop aims to foster interdisciplinary dialogue and innovation at the intersection of AI and education.

Keywords: Epistemics · Decision Making · AI-Supported Education · Multi-modality · Explanability

1 Introduction

The rapid evolution of generative artificial intelligence (GenAI), particularly large language models (LLMs) such as GPT-4, LLaMA, Gemini, and DeepSeek, is fundamentally reshaping the educational landscape [2,18]. These models offer unprecedented capabilities in content generation, personalized learning, and adaptive assessment, heralding a new era of intelligent, inclusive, and student-centered education [3].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 218–224, 2025.
https://doi.org/10.1007/978-3-031-99267-4_27

The integration of GenAI into educational settings enables the creation of dynamic learning materials tailored to individual student needs, fostering engagement and enhancing learning outcomes [1,5]. For instance, LLMs can generate contextually relevant questions and provide real-time feedback, thereby supporting higher-order thinking and problem-solving skills [4,9]. Moreover, GenAI powered intelligent tutoring systems can adapt to learners' progress, offering customized support that aligns with their unique learning trajectories [10].

However, the deployment of GenAI in education also raises critical ethical and practical challenges. Concerns about explainability, fairness, and accountability are paramount [13], as AI-driven tools must be transparent and equitable to gain trust among educators and learners [8,15]. Furthermore, issues such as data privacy, algorithmic bias, and the potential for over-reliance on AI require scrutiny and the development of robust ethical frameworks [6,11,12].

2 Scope

The workshop will feature keynote talks, peer-reviewed papers, and discussions on GenAI's latest research and practical applications in education. The selected papers will be published in the AIED workshop proceedings by Springer. Topics of interest include, but are not limited to:

- Large foundation models for educational content generation and assessment
- Personalized learning and decision-making with generative AI.
- AI-enhanced creativity and problem solving in education.
- Generative AI to support higher-order thinking and creative tasks.
- Agents powered by AI that support adaptivity and collaborative learning.
- Trustworthy and fair AI for education (explainability, fairness, privacy).
- Ethical and security concerns in generative AI for education.
- Generative multi-agent systems for educational decision-making.
- Fine-tuning large models for specific educational contexts and challenges.
- Cases of AI applications in the classroom and the evolving role of teachers in integrating AI into their epistemic practices.

This workshop is designed to foster collaboration and knowledge sharing across fields, promote innovation, and address the ethical and technical challenges posed by generative AI in education. By participating in this workshop, you are not just an attendee, you are an integral part of a community working towards a common goal.

3 Objectives

The overarching goal of the workshop is to provide a multidisciplinary forum for exploring the impact of generative AI on educational theory, practice, and infrastructure. As educational systems worldwide begin to incorporate LLMs

and multimodal generative AI tools, it becomes essential to interrogate their epistemological, pedagogical, ethical, and technical implications [7].

Specifically, the workshop aims to facilitate knowledge exchange on the capabilities and limitations of large foundation models for generating educational content, supporting assessments, and driving personalized instruction. We will We will pursue the following objectives:

- investigate how GenAI influences learners' construction, validation, and application of knowledge, focusing on epistemic agency and critical thinking;
- address issues related to explainability, fairness, transparency, and data privacy to encourage responsible deployment of AI tools in educational contexts;
- promote collaboration among academia, industry, and educational practitioners to develop scalable, evidence-based AI solutions tailored to diverse educational settings;
- examine how educators' roles evolve when AI becomes a co-instructor, co-evaluator, or epistemic agent in the classroom [16];
- provide opportunities for attendees to engage with GenAI technologies through demonstrations, toolkits, and interactive sessions that encourage experimentation, critical analysis, and AI literacy development; and
- identify gaps, open questions, and promising research directions that can lead to new frameworks, algorithms, and applications for AI-supported education.

By targeting these objectives, the workshop seeks to shape a responsible and innovative future for AI in education that centers human values, fosters inclusion, and elevates the quality of learning experiences globally.

4 Motivation

The emergence of GenAI, including LLMs and multimodal systems like GPT-4, LLaMA, Gemini, and DeepSeek, has sparked a paradigm shift in educational technology. These models have opened new avenues for enhancing instructional delivery, generating adaptive assessments, and enabling personalized learning experiences. However, beyond these functional transformations lies a deeper epistemic shift: GenAI is transforming how students construct knowledge, how teachers frame pedagogy, and how decision-making is operationalized in educational systems. This workshop is motivated by the urgent need to explore these epistemic transformations in depth, providing a platform for researchers, educators, and technology developers to collectively understand, critique, and innovate around the use of GenAI in education.

5 Significance to AIED

This workshop is relevant to the AIED community as it bridges foundational advances in AI and core educational values. It seeks to investigate how GenAI can support epistemically rich learning environments, facilitate adaptive teaching strategies, and enable context-aware assessment mechanisms. Meanwhile,

it addresses the sociotechnical challenges accompanying AI integration, such as algorithmic bias, ethical considerations, and educators' evolving role and AI literacy. By foregrounding epistemics and decision-making, the workshop contributes to the theoretical, technical, and applied dimensions of AIED, reinforcing the need for AI systems that are not only intelligent but also pedagogically and ethically sound.

6 Challenges

The integration of GenAI into educational ecosystems presents a series of complex challenges:(a)explainability and transparency—AI-driven decisions should be transparent and interpretable by educators and learners alike [14]. (b)Fairness and privacy–Educational data used to fine-tune or deploy AI models must be handled with care to avoid reinforcing systemic inequities or privacy violation. (c) Aligning AI functionalities with diverse pedagogical goals and epistemic cultures: Educational institutions vary widely in their technological infrastructure, curricular frameworks, and philosophical approaches to learning, making universal solutions difficult to implement. (d) Equitable access–GenAI tools should be accessible to underserved communities and do not exacerbate digital divides. Addressing these challenges requires robust interdisciplinary collaboration and ongoing dialogue among technologists, educators, and policy-makers.

7 Timeline

The workshop follows a structured and time-sensitive preparation phase. The call for submissions remains open until May 22, 2025, after which all papers will undergo a rigorous double-blind review. Notifications of acceptance will be issued by June 1, 2025. The authors of the accepted papers will have two weeks to incorporate feedback and submit their final camera-ready versions by June 14, 2025. The workshop is scheduled on July 26, 2025, as part of the broader AIED 2025 conference. The details about the workshop are available at the website[1].

8 Format

The workshop will be primarily conducted in person, allowing attendees to engage deeply through face-to-face interaction. It will feature a series of invited keynote talks from leading experts, each followed by a Q&A segment to facilitate direct dialogue. Selected papers will be presented through short talks, promoting the dissemination of cutting-edge ideas as much as possible in a concise format. A panel session will bring together diverse voices to discuss overarching issues and future directions for GenAI in education. To engage participants, poster sessions will be held strategically before breaks, and breakout discussions

[1] https://sites.google.com/view/edm-aied-2025/home.

are introduced to encourage intimate and topic-specific exploration. Roundtable discussions will be organized to support early-career researchers and underrepresented scholars in forming close connections with senior academics and practitioners.

9 Activities

The workshop will offer a variety of activities designed to foster collaboration, innovation, and critical inquiry. Keynote talks will highlight the latest theoretical and applied advances in GenAI for education, covering topics such as automated assessment, creativity support, and ethical AI design. Contributed presentations will showcase novel research findings and implementation case studies. Poster sessions will serve as interactive platforms for knowledge exchange, providing participants opportunities to delve into ongoing work and receive feedback. Structured breakout groups will focus on thematic concerns such as personalization, fairness, and epistemic agency, encouraging attendees to co-create insights and future research directions. The panel discussion will offer a comprehensive overview of challenges and opportunities from both academic and industry perspectives. Finally, the event will conclude with closing remarks and the recognition of outstanding contributions through the best paper and poster awards, affirming the workshop's commitment to scholarly excellence and community building.

10 Previous Related Events

Over the past several years, organizers, such as Dr. Zhai, have chaired, presided over, and organized sessions at key academic gatherings such as the American Educational Research Association (AERA), the National Association of Research in Science Teaching (NARST), American Association for the Advancement of Science (AAAS), and the National Council on Measurement in Education (NCME). These events have focused on personalized learning and the ethical implications of AI in education, showcasing the organizers' ability to navigate the intersection of AI and educational research. For example, Dr. Zhai has organized multiple symposia on AI-augmented assessment tools (AERA 2023) and machine learning applications in science assessment (AERA 2020), demonstrating his dedication to both the theoretical and applied aspects of AI in educational contexts [17].

The organizers' leadership roles in these events have positioned them to facilitate meaningful discussions and collaborations among researchers, educators, and industry professionals. For example, Dr. Zhai and colleagues have chaired sessions at NARST on the role of AI in fostering equitable science teaching (2023) and addressing challenges in automated scoring (NCME 2024), reflecting their ongoing commitment to advancing AI-driven innovations that support inclusive and effective learning environments. This extensive involvement in high-impact

conferences also strengthens the organizers' ability to successfully host a workshop that promotes interdisciplinary dialogue and innovation. The workshop will build on the foundation of their prior successful engagements, offering a platform for participants to collaborate and explore in-depth how generative AI can shape the future of education.

Acknowledgment. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305C240010 (PI Zhai) and R305A240356 (PI Liu). The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. Additionally, we acknowledge the use of OpenAI's ChatGPT to polish the language and enhance the clarity of this manuscript.

References

1. Bewersdorff, A., et al.: Taking the next step with generative artificial intelligence: the transformative role of multimodal large language models in science education. Learn. Individ. Differ. **118**, 102601 (2025)
2. FYP24071, C., Runqi, Y., Ruiyi, L., Yuxuan, C., Schnieders, D.: EduLLM: applying large language models in the education industry. AI Educ. 1–17 (2024)
3. Gan, W., Qi, Z., Wu, J., Lin, J.C.W.: Large language models in education: vision and opportunities. In: 2023 IEEE International Conference on Big Data (BigData), pp. 4776–4785. IEEE (2023)
4. Guo, S., Latif, E., Zhou, Y., Huang, X., Zhai, X.: Using generative AI and multi-agents to provide automatic feedback. arXiv preprint arXiv:2411.07407 (2024)
5. Guo, S., Zheng, Y., Zhai, X.: Artificial intelligence in education research during 2013–2023: a review based on bibliometric analysis. Educ. Inf. Technol. **29**(13), 16387–16409 (2024)
6. Latif, E., Zhai, X.: Privacy-preserved automated scoring using federated learning for educational research. arXiv preprint arXiv:2503.11711 (2025)
7. Lee, G., Zhai, X.: Realizing visual question answering for education: GPT-4V as a multimodal AI. In: TechTrends, pp. 1–17 (2025)
8. Łodzikowski, K., Foltz, P.W., Behrens, J.T.: Generative AI and its educational implications. In: Trust and Inclusion in AI-Mediated Education: Where Human Learning Meets Learning Machines, pp. 35–57. Springer (2024)
9. Maity, S., Deroy, A.: The future of learning in the age of generative AI: automated question generation and assessment with large language models. arXiv preprint arXiv:2410.09576 (2024)
10. Maity, S., Deroy, A.: Generative AI and its impact on personalized intelligent tutoring systems. arXiv preprint arXiv:2410.10650 (2024)
11. Nguyen, A., Ngo, H.N., Hong, Y., Dang, B., Nguyen, B.: Ethical principles for artificial intelligence in education. Educ. Inf. Technol. **28**(4), 4221–4241 (2023)
12. Raji, I.D., Scheuerman, M.K., Amironesei, R.: You can't sit with us: exclusionary pedagogy in AI ethics education. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 515–525 (2021)
13. Ts, A., Biswas, G.: Relating students cognitive processes and learner-centered emotions: an advanced deep learning approach. In: Proceedings of the 26th International Conference on Multimodal Interaction, pp. 575–584 (2024)

14. Wu, X., Saraf, P.P., Lee, G., Latif, E., Liu, N., Zhai, X.: Unveiling scoring processes: dissecting the differences between LLMs and human graders in automatic scoring, pp. 1–16. Technology, Knowledge and Learning (2025)
15. Wu, X., et al.: Usable XAI: 10 strategies towards exploiting explainability in the LLM era. arXiv preprint arXiv:2403.08946 (2024)
16. Zhai, X.: Transforming teachers' roles and agencies in the era of generative AI: perceptions, acceptance, knowledge, and practices. J. Sci. Educ. Technol. 1–11 (2024)
17. Zhai, X., Krajcik, J.: Uses of Artificial Intelligence in STEM Education. Oxford University Press (2024)
18. Zhai, X., Yin, Y.: Using educative learning progression to support novice science teachers' lesson plan critiques. Int. J. Sci. Educ. **46**(13), 1281–1310 (2024)

D-SAIL - Transformative Curriculum Design: Digitalisation, Sustainability, and AI Literacy for 21st Century Learning

Martin Ruskov[1(✉)], Dimitri Ognibene[2,3], Davinia Hernández-Leo[4], Davide Taibi[5], Yannis Dimitriadis[6], Giovanni Fulantelli[5], Uwe Handmann[7], George K. Zarifis[8], Yan Wu[9], and Paolo Maria Ferri[2]

[1] University of Milan, 20100 Milan, Italy
martin.ruskov@unimi.it
[2] University of Milan Bicocca, 20100 Milan, Italy
{dimitri.ognibene,paolo.ferri}@unimib.it
[3] National Research Council - Institute for Cognitive Sciences and Technologies, 00196 Rome, Italy
[4] Pompeu Fabra University, 08002 Barcelona, Spain
davinia.hernandez-leo@upf.edu
[5] National Research Council - Institute for Educational Technology, 90146 Palermo, Italy
{davide.taibi,giovanni.fulantelli}@itd.cnr.it
[6] University of Valladolid, 47011 Valladolid, Spain
yannis@tel.uva.es
[7] Hochschule Ruhr West, 45407 Mülheim an der Ruhr, Germany
uwe.handmann@hs-ruhrwest.de
[8] Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece
gzarifis@edlit.auth.gr
[9] Agency for Science, Technology and Research, Singapore 138632, Singapore
wuy@i2r.a-star.edu.sg

Abstract. This half-day workshop invites educators, researchers, practitioners, and other stakeholders to reimagine traditional curricula by integrating AI-driven insights, sustainability principles, and digitalisation. The rapidly evolving technological landscape around teaching and learning calls for innovative approaches to foster adaptive, inclusive, and future-ready learning experiences. The workshop will explore how generative AI can enhance smart learning environments and how comprehensive data analysis can inform effective curriculum development. Current efforts lay the groundwork for embedding data analysis and AI methodologies across diverse higher education disciplines. However, concerns have been raised about the unequal access to these technologies. These directions can provide valuable insights into creating transformative educational experiences that not only leverage cutting-edge digital tools but also embed sustainability and ethical practices into curriculum design. Multidisciplinary research could reveal limitations of current approaches

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 225–232, 2025.
https://doi.org/10.1007/978-3-031-99267-4_28

and propose improvements and alternatives. A non-exhaustive list of topics of interest is innovative pedagogical frameworks, sustainability and digitalisation in education, human-AI collaboration, ethical and societal considerations, and others. The workshop will provide a platform to share research findings, discuss innovative pedagogical frameworks, and explore interdisciplinary strategies that prepare students for the demands of the 21st century.

Keywords: transformative curriculum design · digitalisation · sustainability · AI literacy

1 Introduction

Generative AI has become a continuous source of change in our society. Inevitably, this has an enormous impact on teaching and learning, with adoption happening at a diverging pace [4,7]. While some excitedly rush to embrace opportunities [3,6,8,9], others warn of threats and challenges related to unequal access, model hallucinations, biases in models, environmental cost, etc. [1,2]. In the rapidly evolving technological landscape, both innovative approaches and reflections on the state of affairs are needed to foster an adaptive, inclusive, and future-ready educational ecosystem. Generative AI has grown far beyond the traditional boundaries of computer and data sciences and has become relevant to all of our society [3,5]. This is why an interdisciplinary and holistic approach towards developing AI literacy is crucial for sustainable education and learning. This is particularly so considering the broad spectrum of methodologies in education from formal sciences to humanities, from cognitive to physical, from coaching to self-directed learning.

This workshop invites educators, researchers, practitioners and other stakeholders to reimagine traditional curricula across the disciplines by integrating AI-driven insights, sustainability principles, and digitalisation. The workshop will explore how generative AI can enhance smart learning environments and how comprehensive data analysis can inform effective curriculum development. Ongoing research is demonstrating the potential of human-centered, AI-enhanced smart learning environments that support hybrid learning scenarios by overcoming conventional limitations. Moreover efforts are being made to lay the groundwork for embedding data analysis and AI methodologies across diverse educational disciplines. When considered in combination with potential risks, these directions can provide valuable insights into creating transformative educational experiences that not only leverage cutting-edge digital tools but also embed sustainability and ethical practices into curriculum design. Multidisciplinary research could reveal limitations of current approaches and propose improvements and alternatives. The workshop will provide a platform to share research findings, discuss innovative pedagogical frameworks, and explore interdisciplinary strategies that prepare students and learners for the demands of the 21st century.

2 Call for Papers

We invite submissions of papers that examine and expand upon innovative strategies for curriculum transformation at the intersection of AI, sustainability, and digitalisation. We welcome original research, case studies, conceptual frameworks, and best practice reports. Real-world research, such as in-the-wild studies, design-based research, and analysis of public data are of particular interest.

Relevant topics include, but are not limited to:

Innovative Pedagogical Frameworks: Approaches that integrate AI-driven insights and data analytics into curriculum design to create adaptive and personalized learning environments.

Sustainability and Digitalisation in Education: Strategies for embedding sustainability principles and digital tools into curricula, ensuring that educational practices are both future-proof and environmentally responsible.

Human-AI Collaboration: Case studies and research on effective models for human-AI collaboration in the design and delivery of learning experiences, including insights from generative AI applications. Interdisciplinary Curriculum Integration: Approaches for leveraging data analysis and AI across various academic disciplines.

Ethical and Societal Considerations: Discussions on the ethical challenges and societal implications of integrating advanced digital technologies into educational settings, with a focus on inclusivity and equitable access.

Submissions should clearly articulate both theoretical and practical dimensions of curriculum transformation, with an emphasis on real-world applications, challenges, and future directions. Authors are encouraged to discuss the implications for teaching, learning, and academic policy in a digitally transformed educational landscape.

Submissions should be anonymized, as part of a double-blind review process. Authors of accepted papers will be allocated slots to present their contributions.

2.1 Submission Guidelines

Contributions should present original research and clearly articulate both theoretical and practical dimensions of curriculum transformation, with an emphasis on real-world applications, challenges, and future directions. Authors are encouraged to discuss the implications for teaching, learning, and academic policy in a digitally transformed educational landscape.

Accepted paper formats:

- Full papers (10-12 pages) - original research with detailed analysis and conclusive findings

- Short papers (6-9 pages) - work in progress, early results feasibility studies or reproducibility studies.
- Project and position papers (5 pages) - new visions or initiatives, building on rigorous previous research.

We envision 8 presentations (invited and submitted) of interest to educators, researchers, and practitioners, curriculum designers and policymakers. Submissions should be anonymized, as part of a double-blind review process. All submissions will be evaluated by at least two experienced reviewers and detailed feedback will be provided to each submission. Authors of accepted papers will be asked to present their contributions orally and with slides. Each paper will be allocated 15 min (short) or 20 min (long) with 5 more minutes for questions and discussion.

Proceedings will be published with CEUR-WS, will be assigned an URN and indexed by SCOPUS, Web of Science, and other major indexes. Authors are asked to follow the one-column CEURART format for their submissions.

2.2 Important Dates

All dates are in the Anywhere-on-Earth time zone.

- Abstract submission deadline: 19 May 2025
- Paper submission deadline: 26 May 2025
- Acceptance notification: 20 June 2025
- Camera ready submissions: 1 July 2025
- Workshop: 22 July 2025

Join us in shaping the future of higher education by transforming curricula to meet the evolving demands of a digital, sustainable, and AI-enhanced world.

3 Oranizers

Martin Ruskov is a researcher at the Department of Languages, literatures, cultures and mediations at the University of Milan. He has stated his research in the field of game-based learning, working on numerous international research projects with threshold concepts, variation theory and design-based research at University College London, Zentrum für grafische Datenverarbeitung, Darmstadt and Sofia University. He has also spent a period outside of academia working on mobile-assisted language learning, offender rehabilitation and gamification. His research interests include awareness development, life-long learning and intangible cultural heritage.

Dimitri Ognibene is an Associate Professor at the University of Milano-Bicocca and an Associate Researcher at ISTC CNR in Rome, whose multidisciplinary work spans artificial intelligence, robotics, and cognitive neuroscience. His distinguished career includes roles at Universitat Pompeu Fabra,

King's College London, and Imperial College London, and most recently he has been serving as Visiting Professor in the Department of Education at the University of Tokyo. His research bridges theory and practice—most notably in exploring the interplay between attention and learning in computational models. While his Ph.D. work underscores the importance of this interaction, his varied publication record suggests that exceptions are indeed possible.

Davinia Hernández-Leo is a Full Professor at the Universitat Pompeu Fabra's Department of Engineering and a Serra Húnter and Icrea Academia Fellow. Her multidisciplinary work bridges human-computer interaction, network applications, and learning sciences, positioning her as a leader in technology-enhanced education. As head of the Interactive and Distributed Technologies for Education (TIDE) research group, she has made significant contributions to learning design and computer-supported collaborative learning. She is currently engaged in proposing data-driven and AI-supported methods to improve learning design and smart learning environments, emphasizing human rights and ethics.

Davide Taibi is a Senior Researcher at the Institute for Educational Technology of the National Research Council of Italy (CNR-ITD). His research primarily focuses on the application of innovative technologies to support education at various levels, covering areas such as Mobile Learning, Semantic Web and Linked Data for e-learning, standards for educational processes design, Open Education, Learning Analytics, Augmented and Virtual Reality for education, and Artificial Intelligence in Education. He has coordinated two EU funded projects in the field of Data Literacy: DATALIT (Data Literacy at the interface of higher education and business), and DEDALUS (DEveloping DAta Literacy courses for University Students). Currently, he is leading two EU funded project focused on Artificial Intelligence in Education named SMERALD (SMEs Raising Awareness and Learning on Digital data, data analysis and artificial intelligence) and IDEAL (Integrating Data Analysis and AI in Learning experiences). He is also a contract professor at the University of Palermo.

Yannis Dimitriadis is Full Professor of Telematics Engineering and ex Dean of the Doctoral School, University of Valladolid, Spain. He is also the coordinator of the GSIC/EMIC research group, an inter-disciplinary group, integrating over 20 researchers and practitioners from the field of Information and Communications Technologies and Pedagogy. His recent research work has focused on learning analytics and smart learning environments, alignment of learning design and learning analytics, design patterns, conceptual and technological support to the orchestration of computer-supported collaborative learning processes, active pedagogies at scale, and across-spaces (Web, 3D worlds and augmented reality) learning. He has participated in more than 50 competitive research projects on technology-enhanced learning, co-authored more than 100 journal papers and 215 conference papers (including several papers at LAK), and organized several workshops and symposia, at ECTEL, CSCL, ISLS, etc. Dr. Dimitriadis is also a senior member of IEEE, member of ISLS.

Giovanni Fulantelli is Senior Researcher at the Institute for Educational Technology of the National Research Council of Italy (CNR), and contract professor at the LUMSA University, where he holds a course in Digital technologies for learning. In 2005 he was awarded as one of the excellent researchers at the CNR. His research interests are inspired, on the one hand, by the open education movement, and the principle that technologies must and can guarantee universal access to quality education; on the other hand, by the awareness that the rapid diffusion of social media and AI-based solutions requires special attention to avoid compromising the potential for educational development and personal growth offered by digital technologies.

Uwe Handmann is a distinguished academic and industry expert whose multifaceted work spans neuroinformatics, artificial intelligence, and sensor data fusion. He is a Professor of Neuroinformatics at Hochschule Ruhr West University of Applied Sciences, He also directs the Computer Science Institute and leads multiple innovation centers focused on digitalization and the twin transition towards a circular economy. His extensive career includes leadership roles in both academia and industry, from heading research projects on intelligent systems and image processing to serving as Managing Director at iQServ GmbH. He also enriches international academic collaboration as a Visiting Professor at Babeş-Bolyai University in Cluj, Romania. While his early research laid a strong foundation in AI and computational models, his later work reflects a broad engagement with cutting-edge applications and digital transformation, highlighting that even focused research has a strong practical relevance.

George K. Zarifis is Professor of Continuing Education in the Faculty of Philosophy, Department of Education at the Aristotle University of Thessaloniki. His research interests focus on adult educators' and teachers' training and professionalisation, university continuing education, curriculum design and learning outcomes, and comparative examination of adult learning and vocational education and training policies and practices in Europe. He is the convener of the ESREA's Research Network on the Professional Development of Adult Educators (ReNAdET). He publishes, edits and co-authors in the area of adult and continuing education, and participates in a large number of national and European research projects in the same field.

Yan Wu is the Deputy Head of the Robotics and Autonomous Systems Department at A*STAR Institute for Infocomm Research, where he also works as a Principal Scientist and Leader of the Manipulation Group. Since Sept 2023, he is concurrently the Co-Lead (Technology) at the Robotics Horizontal Technology Coordinating Office at Agency for Science, Technology and Research, Singapore. Yan is the current Chair of the IEEE Systems, Man and Cybernetics Society, Singapore Chapter and Member of the IEEE Robotics and Automation Society Technical Committees on Cognitive Robotics, Haptics and Neuro-Robotics Systems. He is also the current President of the Pattern Recognition and Machine Intelligence Association. He serves or has served as an Associate Editor at various editorial boards, such as IEEE Transactions on Automation Science, IEEE

Robotics and Automation Society Conference Editorial Board, IEEE Intelligent Transport Systems Society Conference Editorial Board and Frontier in Robotics and AI. Yan is a Senior Member of the IEEE. His research interests include dexterous manipulation, embodied robot learning and interaction, and service and assistive robotics.

Paolo Maria Ferri is a Full Professor of Teaching and Special Pedagogy at the University of Milan Bicocca. He coordinates research in digital learning environments and "digitally augmented" pedagogies within the Department of Human Sciences "Riccardo Massa." His work focuses on integrating emerging technologies, such as AI, into formal education, with particular attention to 0-10 children and primary school. He has supervised doctoral students and contributed to national guidelines for MOOCs on behalf of CRUI. Ferri has led several funded projects under Horizon 2020 and the Italian Ministry of Education (Miur). He regularly consults for leading publishing groups, designing digital teaching content and training programs for educators. His publications include "Nativi Digitali" (Bruno Mondadori, 2011) and "I Nuovi Bambini" (Rizzoli, 2014), among others. He was a founding member of SIREM and serves on editorial boards of academic journals such as TD and Je-LKS. Ferri's ongoing research investigates the interplay between digital media appropriation, teacher professional development, and e-learning, with a particular focus on the use of Ai in Schools and Universities.

3.1 Program Committee Members

- Ishari Amarasinghe, Open University of Catalonia, Spain
- Albena Antonova, Sofia University, Bulgaria
- Gabriella Casalino, University of Bari, Italy
- Alessandro Ciasullo, University Federico II of Naples, Italy
- Lukas Erle, Hochschule Ruhr West University of Applied Sciences, Germany
- Raffaele Di Fuccio, Pegaso University, Italy
- Cansu Koyuturk, University of Milan Bicocca, Italy
- Carla Limongelli, Roma Tre University, Italy
- Nicola Orio, University of Padova, Italy
- Maria Perifanou, University of Macedonia, Greece
- Marco Rondonotti, e-campus University, Italy
- Daniele Schicchi, ITD CNR, Italy
- Giulia Sironi, University of Milan Bicocca, Italy
- Marco Temperini, Sapienza University, Italy
- Emily Theophilou, Pompeu Fabra University, Spain
- Ricardo Torres Kompen, Ramon Llull University, Spain
- Daniel Zilio, University of Padua, Italy

References

1. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. ACM, Virtual Event Canada (2021). https://doi.org/10.1145/3442188.3445922
2. Chen, L., Zaharia, M., Zou, J.: How is ChatGPT's behavior changing over time? Harvard Data Sci. Rev. **6**(2) (2024). https://doi.org/10.1162/99608f92.5317da47
3. Cuscito, M., Ferrara, A., Ruskov, M.: How BERT speaks Shakespearean English? evaluating historical bias in contextual language models. In: Proceedings of the 3rd Workshop on Artificial Intelligence for Cultural Heritage, pp. 14–21. CEUR-WS, Bolzano, Italy (2024). https://ceur-ws.org/Vol-3865/02_paper.pdf
4. Giannakos, M., et al.: The promise and challenges of generative AI in education. Behav. Inf. Technol. 1–27 (2024). https://doi.org/10.1080/0144929X.2024.2394886
5. Humlum, A., Vestergaard, E.: Large language models. Small Labor Market Effects (2025). https://doi.org/10.2139/ssrn.5219933
6. Kambhampati, S., Stechly, K., Valmeekam, K.: (How) Do reasoning models reason? Ann. New York Acad. Sci. nyas.15339 (2025). https://doi.org/10.1111/nyas.15339
7. Noroozi, O., Soleimani, S., Farrokhnia, M., Banihashem, S.K.: Generative AI in education: pedagogical, theoretical, and methodological perspectives. Int. J. Technol. Educ. **7**(3), 373–385 (2024). https://doi.org/10.46328/ijte.845
8. Ognibene, D., et al.: Use me wisely: AI-driven assessment for LLM prompting skills development. Educ. Technol. Soc. **28**(3), 184–201 (2025). https://doi.org/10.30191/ETS.202507_28(3).SP12
9. Ruskov, M.: Who and how: using sentence-level NLP to evaluate idea completeness. In: Wang, N., Rebolledo-Mendez, G., Dimitrova, V., Matsuda, N., Santos, O.C. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, pp. 284–289. Communications in Computer and Information Science, Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-36336-8_44

Second Workshop on the Automatic Evaluation of Learning and Assessment Content

Luca Benedetto[1(✉)], Diana Galvan-Sosa[1], George Dueñas[2], Shiva Taslimipoor[1], Gabrielle Gaudeau[1], Andrew Caines[1], Anastassia Loukina[3], and Torsten Zesch[4]

[1] ALTA Institute, Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
{luca.benedetto,diana.galvan-sosa,shiva.taslimipoor,gabrielle.gaudeau, andrew.caines}@cl.cam.ac.uk
[2] National Pedagogical University, Bogota, Colombia
geduenasl@upn.edu.co
[3] Grammarly Inc., San Francisco, USA
anastassia.loukina@grammarly.com
[4] FernUniversität, Hagen, Germany
torsten.zesch@fernuni-hagen.de

Abstract. Evaluation of learning and assessment content has always been a crucial task in the educational domain, but traditional approaches based on human feedback are not always usable in modern educational settings. Indeed, the advent of machine learning models, in particular Large Language Models (LLMs), enabled to quickly and automatically generate large quantities of data, making human evaluation unfeasible. Similarly, Massive Open Online Courses (MOOCs) have numbers of attending students so large that it is unsustainable to manually provide feedback to all of them. Thus, the need for accurate and automated techniques for evaluating educational content – e.g., questions, hints, and feedback – became pressing. Building on the success of the First Workshop on the Automatic Evaluation of Learning and Assessment Content, which was held at AIED 2024, this workshop aims to attract professionals from both academia and industry, and to offer an opportunity to discuss which are the common challenges in evaluating learning and assessment content and tools in education.

Keywords: Automated Evaluation · Natural Language Processing · Language Models · Educational NLP

1 Workshop Chairs

1.1 General Chair

– **Luca Benedetto** is a research associate at the University of Cambridge, working in the ALTA institute (Automated Language Teaching and Assessment) and NLIP group, and holds a doctoral degree from Politecnico di

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 233–240, 2025.
https://doi.org/10.1007/978-3-031-99267-4_29

Milano (Italy). His research primarily focuses on the applications of Natural Language Processing (NLP) techniques to education, such as automated evaluation and generation of content, and the study of the differences and similarities between knowledge acquisition in human learners and knowledge of Large Language Models.

1.2 Program Chairs

- **Diana Galvan-Sosa** is a research associate at the University of Cambridge, working in the ALTA institute (Automated Language Teaching and Assessment) and NLIP group, and holds a doctoral degree from Tohoku University (Japan). Her research interests include Natural Language Processing, particularly knowledge acquisition and applications of NLP to education such as feedback generation for L2 (second language) learners.
- **George Dueñas** is a PhD student in the Interinstitutional Doctorate in Education at the National Pedagogical University (Colombia), passionate about using Large Language Models to improve educational experiences. His research explores the intersection of language, education, and computing, focusing on areas such as automatic prediction of item difficulty, automated essay scoring and reading comprehension.

1.3 Publicity Chairs

- **Shiva Taslimipoor** is a senior research associate in the NLIP group and ALTA institute at the University of Cambridge, and holds a doctoral degree from University of Wolverhampton. Her research lies in the intersection of NLP and machine learning and at ALTA she works on machine learning technologies for language teaching and assessment. Following her work on Cloze item generation, her research is now more focused on educational content creation and evaluation.
- **Gabrielle Gaudeau** is a 2nd-year PhD student in the NLIP group of the Department of Computer Science and Technology at the University of Cambridge. Her current research focuses on automating written assessment and feedback generation for second-language learners, supported by the Institute for Automated Language Teaching and Assessment (ALTA) and Cambridge University Press & Assessment.

1.4 Advisors

- **Andrew Caines** is a senior research associate in the NLIP Group & ALTA Institute based in the Computer Laboratory at the University of Cambridge, U.K. His research relates to automated assessment and feedback for learners of English as a second language, as well as the personalisation of education

technology and evaluation of the capabilities of large language models. He has previously organised a shared task at the NLP4CALL workshop, the IVACS conference 2024, one-day workshops for early careers researchers, and departmental seminar series.

- **Anastassia Loukina**, is an Engineering Manager at Grammarly, Inc. She leads a cross-functional team responsible for building tools to monitor the quality of Grammarly AI systems. Before joining Grammarly, Anastassia was a senior research scientist in the Research and Development division at Educational Testing Service (ETS) in Princeton, NJ, where she worked on improving the validity, reliability, and fairness of speech-based educational applications. She has published over 60 papers and book chapters, holds several patents, and frequently attends international conferences and workshops. Before joining ETS, Anastassia worked at the Oxford University Phonetics Laboratory, where she conducted research and taught graduate and undergraduate courses in linguistics and phonetics. She holds an MPhil and DPhil from the University of Oxford.
- **Torsten Zesch** is a full professor of Computational Linguistics at the Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA) at FernUniversität in Hagen, Germany. He holds a doctoral degree in computer science from Technische Universität Darmstadt and was the president of the German Society for Computational Linguistics and Language Technology (GSCL) from 2017 to 2023. His main research interests are in educational natural language processing, in particular the ways in which teaching and learning processes can be supported by language technology and AI. For this purpose, he develops methods for the automatic analysis of textual and multimodal language data, with a focus on robust and explainable models.

2 Content, Themes, Relevance to the AIED Community

Evaluation of learning and assessment content has always been very important in the educational domain, regardless of the specific setting: K-12, MOOCs, language certification exams, and learning apps (to name some settings frequently studied in the literature) all require that the learning and assessment content is evaluated to ensure high quality and good learning outcomes. In online settings and, in general, settings that involve a lot of students or that require large libraries (e.g., of questions), it is unfeasible to exclusively rely on subject matter experts to perform such evaluations. Thus, there is the need for accurate and reliable techniques for automatically evaluating learning and assessment contents, either with human-in-the-loop approaches or in fully automated ones, depending on the needs and constraints of each use case.

Assessment content, such as questions and exams, is commonly evaluated both with traditional approaches based on students' responses – e.g., Item Response Theory [12] – and more modern approaches based on machine learning and Natural Language Processing [1,3,5]. More recently, some research also

attempted to perform this evaluation by simulating the responses of students with AI models and performing virtual pretesting [4,15,16,22]. Another challenge in automated evaluation is related to distractors (i.e., the plausible incorrect options in multiple choice questions): they are commonly evaluated with manual labelling [7,9,10], since automatic evaluation approaches proposed so far have some limitations [18,21], but manual evaluation cannot scale to production settings in, for example, online learning platforms with large libraries of questions. On the other hand, the evaluation of learning content – such as single lectures, and whole courses and curricula – still relies heavily on experts from the educational domain, but could be potentially automated, at least partially and with a human-in-the-loop approach (e.g., flagging pre-requisites which are not satisfied by the previous lectures in a course). Although not strictly "content", Virtual Teaching Assistants (VTAs) also have to be evaluated, and this is particularly relevant with modern LLM-based VTAs [2,20], since biases and behaviours from other tasks and domains could be unwillingly transferred to the educational domain. Indeed, older models were often rule-based, thus more limited and less powerful but more controllable (and not subject to hallucinations) [11], while more modern assistants are more powerful and flexible but need to be carefully evaluated. LLMs as a whole need to be evaluated from a pedagogical perspective – this is usually referred to as *pedagogical alignment* of LLMs [17,19] – since they can be and are being used for a variety of educational tasks, in addition to serving as virtual tutors, as shown in previous research (from feedback generation and automated assessment to question and content generation [8,13,14]). As an example, generated content must align with the learning objectives of the specific course (or exam) where it is being used, and use language of a complexity level suitable for the target students. Importantly, this evaluation must consider not only the educational requirements of the generated content, which is the primary focus of pedagogical alignment, but also the biases that might emerge from the generation models, to ensure EDIB (Equity, Diversity, Inclusion, & Belonging) appropriateness, and the factual accuracy of the generated text.

Building on the success of the First Workshop on the Automatic Evaluation of Learning and Assessment Content [6], which was held at AIED 2024 and attracted more than 80 participants, this workshop will focus on approaches for automatically evaluating learning and assessment content, offering an opportunity to discuss common challenges, share best practices, and promising new research directions. Topics of interests include but are not limited to:

- Question evaluation (e.g., in terms of the pedagogical criteria listed above: alignment to the learning objectives, factual accuracy, language level, cognitive validity, etc.).
- Estimation of question statistics (e.g., difficulty, discrimination, response time, etc.).
- Evaluation of distractors in Multiple Choice Questions.
- Evaluation of reading passages in reading comprehension questions.
- Evaluation of lectures and course material.

- Evaluation of learning paths (e.g., in terms of prerequisites and topics taught before a specific exam).
- Evaluation of educational recommendation systems (e.g., personalised curricula).
- Evaluation of hints and scaffolding questions, as well as their adaptation to different students.
- Evaluation of automatically generated feedback provided to students.
- Evaluation of techniques for automated scoring.
- Evaluation of pedagogical alignment of LLMs.
- Evaluation of the ethical implications of using open-weight and commercial LLMs in education.
- Evaluation of bias in educational content and LLM outputs.

Human-in-the-loop approaches are welcome, provided that there is also an automated component in the evaluation and there is a focus on the scalability of the proposed approach. Papers on generation are also very welcome, as long as there is an extensive focus on the evaluation step.

As in the previous year, we expect this workshop to attract professionals from both industry and academia, in terms of both submissions and attendees, and to offer an opportunity to discuss the common challenges in evaluating learning and assessment content in education. From the submitted papers and through the interactive and QA sessions which will take place during the workshop, we aim at discussing some guidelines and best practices for the evaluation of educational content, which will be very relevant for the AIED community.

2.1 Length, Format, and Activities

The workshop will be scheduled as a **full-day** workshop, with a format in line with the proposal below (the format might be subject to changes).

Proposed Format

- Introduction by the organisers;
- Talks from the invited speaker(s);
- Paper presentations;
- Hand-on interactive session;
- Discussions and concluding remarks.

Invited Speakers. The workshop will feature an invited talk by Dr. Guanliang Chen, Senior Lecturer in the Faculty of Information Technology at Monash University. A second invited speaker might be announced at a later date.

Paper Format and Paper Presentations. The call for papers is open to both short (5 pages, excluding references) and long papers (10 pages, excluding references). Additionally, we welcome submissions of extended abstracts (2 pages) to provide a forum for ongoing research and project proposals. The proceedings volume will be submitted for publication to *CEUR Workshop Proceedings*, including only long and short papers. If allowed by the venue, all papers will be presented in the form of posters, to maximise the chance of interactions and discussions among workshop participants.

Hands-On Interactive Session. This year, the workshop will also include a hands-on interactive session, curated and hosted by the workshop organisers, to give the participants the possibility of experimenting with modern AI techniques in the context of content evaluation. The session is planned to focus on common (reference and referenceless) metrics used in Natural Language Generation (NLG). These metrics are a valuable tool to evaluate any text generation systems, including LLMs. The activities will be planned with a broad audience in mind, requiring no prior experience with NLP.

Discussion and Concluding Remarks. The discussion session held at the end of the workshop will focus on the main takeaways and general rules of thumb which can be helpful in evaluating learning and assessment content.

3 Timeline

- Launch of the workshop website and first call for papers: between March 21, 2024 and March 28, 2025
- *Workshop website published on the conference website: March 28, 2025*
- Submission deadline: May 25, 2025
- Reviewing phase: from May 27 to June 14, 2025
- Notification of acceptance: June 15, 2025
- *Workshop: 26 July 2025*

Expected Target Audience: With a focus on the automated evaluation of educational content, both manually and automatically generated, this workshop can be of interest to both industry and academia, as it will shed some light on possible approaches to automatically evaluate learning content. In this sense, it can appeal to most of the AIED community that will be attending the conference. Last year, we welcomed more than 80 participants (according to the registration form), and we hope to hold an even larger workshop this year.

4 Program Committee Members

Several distinguished researchers from both academia and industry already agreed to serve as members of the program committee:

- Terry Payne, University of Liverpool (United Kingdom)
- Sebastian Gombert, DIPF — Leibniz Institute for Research and Information in Education (Germany)
- Russell Moore, University of Cambridge (United Kingdom)
- Ali Al-Zawqari, Vrije Universiteit Brussel (Belgium)
- Anna Winklerová, Masaryk University (Czechia)
- David de Fitero Domínguez, University of Alcalá (Spain)
- Benny Johnson, VitalSource Technologies (United States)
- Luis Gabriel Moreno Sandoval, Pontificia Universidad Javeriana (Colombia)
- Ulrike Pado, Hochschule fuer Technik Stuttgart (Germany)
- Radek Pelánek, Masaryk University (Czechia)
- Nicolas Ballier, Université Paris Cité (France)
- Yoav Bergner, New York University (United States)
- Jesús Ángel del Hoyo Gabaldón, Universidad de Alcalá (Spain)
- Samah Alkhuzaey, Umm Al-qura University (Saudi Arabia)
- Rositsa Ivanova, University of St. Gallen (Switzerland)
- Andrea Cappelli, QA Ltd. (Switzerland)

References

1. AlKhuzaey, S., Grasso, F., Payne, T.R., Tamma, V.: Text-based question difficulty prediction: a systematic review of automatic approaches. Int. J. Artif. Intell. Educ. 1–53 (2023)
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)
3. Benedetto, L.: A quantitative study of NLP approaches to question difficulty estimation, pp. 428–434 (2023)
4. Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., Buttery, P.: Using LLMs to simulate students' responses to exam questions. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 11351–11368 (2024)
5. Benedetto, L., et al.: A survey on recent approaches to question difficulty estimation from text. ACM Comput. Surv. (CSUR) (2022)
6. Benedetto, L., et al.: Workshop on automatic evaluation of learning and assessment content. In: International Conference on Artificial Intelligence in Education, pp. 473–477. Springer (2024)
7. Bitew, S.K., Deleu, J., Develder, C., Demeester, T.: Distractor generation for multiple-choice questions with predictive prompting and large language models. arXiv preprint arXiv:2307.16338 (2023)
8. Caines, A., et al.: On the application of large language models for language teaching and assessment technology (2023)

9. Chamberlain, D.J., Jeter, R.: Creating diagnostic assessments: automated distractor generation with integrity. J. Assess. High. Educ. **1**(1), 30–49 (2020)
10. Ghanem, B., Fyshe, A.: Disto: evaluating textual distractors for multi-choice questions using negative sampling based approach. arXiv preprint arXiv:2304.04881 (2023)
11. Goel, A.K., Polepeddi, L.: Jill watson: a virtual teaching assistant for online education. In: Learning Engineering for Online Education, pp. 120–143. Routledge (2018)
12. Hambleton, R.K., Swaminathan, H.: Item Response Theory: Principles and Applications. Springer Science & Business Media (2013)
13. Jeon, J., Lee, S.: Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. Educ. Inf. Technol. **28**(12), 1–20 (2023)
14. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
15. Maeda, H.: Field-testing multiple-choice questions with AI examinees: English grammar items. Educ. Psychol. Meas. **85**(2), 00131644241281053 (2024)
16. Park, J.W., Park, S.J., Won, H.S., Kim, K.M.: Large language models are students at various levels: zero-shot question difficulty estimation. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 8157–8177 (2024)
17. Razafinirina, M.A., Dimbisoa, W.G., Mahatody, T.: Pedagogical alignment of large language models (LLM) for personalized learning: a survey, trends and challenges. J. Intell. Learn. Syst. Appl. **16**(4), 448–480 (2024)
18. Rodriguez-Torrealba, R., Garcia-Lopez, E., Garcia-Cabot, A.: End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. Expert Syst. Appl. **208**, 118258 (2022)
19. Sonkar, S., Ni, K., Chaudhary, S., Baraniuk, R.G.: Pedagogical alignment of large language models. arXiv preprint arXiv:2402.05000 (2024)
20. 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)
21. Taslimipoor, S., Benedetto, L., Felice, M., Buttery, P.: Distractor generation using generative and discriminative capabilities of transformer-based models. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 5052–5063 (2024)
22. Uto, M., Tomikawa, Y., Suzuki, A.: Question difficulty prediction based on virtual test-takers and item response theory. In: Workshop on Automatic Evaluation of Learning and Assessment Content (2024)

Human-AI Collaboration and Culture For Equity and Inclusion (Half-Day Workshop)

Nia Nixon[1], Srecko Joksimovic[2], Andrew Zamecnik[2], Mohammad Amin Samadi[1(✉)], Jaeyoon Choi[1], Seehee Park[1], Valéry Psyché[3], Mohamed El Habib Laraba[5], and Emmanuel G. Blanchard[4]

[1] University of California, Irvine, CA 92618, USA
{dowelln,masamadi,jaeyoon.choi,seeheep}@uci.edu
[2] University of South Australia, Adelaide 5001, Australia
{srecko.joksimovic,andrew.zamecnik}@unisa.edu.au
[3] Department of Education, TELUQ University, Quebec, Canada
valery.psyche@teluq.ca
[4] Le Mans University, Le Mans, France
emmanuel.blanchard@univ-lemans.fr
[5] University of Abdelhamid Mehri Constantine 2, El Khroub, Algeria
habib.laraba@univ-constantine2.dz

Abstract. As artificial intelligence (AI) moves beyond being a passive tool toward becoming an active teammate in learning environments, this workshop explores the opportunities and challenges of Human-AI teaming for fostering inclusive collaboration in education. Drawing from the Machines as Teammates (MaT) framework, we will discuss how AI can be designed not just to support learning tasks, but to function as an adaptive, equity-focused teammate that enhances group dynamics, mitigates bias, and fosters more equitable participation in collaborative problem-solving. The workshop will bring together AI researchers, educators, EdTech developers, and policymakers to examine key questions: What role should AI play in collaborative learning? How can AI teammates promote inclusion, especially in STEM fields where representation gaps persist? What ethical considerations must be addressed to ensure AI enhances rather than disrupts team dynamics? Through expert presentations, a panel discussion, and interactive breakout sessions, participants will explore design principles for AI teammates that support rather than replace human collaboration, with a focus on real-world educational applications. This workshop is designed for those interested in the intersection of AI, collaboration, and equity in education, and aims to drive forward a research and policy agenda on how AI teammates can shape the future of inclusive learning environments.

Keywords: Human-AI teaming · Inclusion · Collaboration · Culture · Equity

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 241–248, 2025.
https://doi.org/10.1007/978-3-031-99267-4_30

1 Organizers and Contact Information

1.1 Workshop Chairs

Dr. Nia Nixon, University of California, Irvine, USA

Email dowelln@uci.edu
Position Assistant Professor, School of Education & Department of Cognitive Science
Bio Dr. Nia Nixon's is a leading researcher in learning analytics and artificial intelligence in education. She also holds a Distinguished International Professorship at the Faculty of Economic and Social Sciences, Hector Research Institute for Education Sciences and Psychology at the University of Tübingen. Her research spans learning analytics, AI in education, and inclusive STEM collaboration, with a focus on socio-cognitive dynamics in human and AI teams.

Dr. Srecko Joksimovic, University of South Australia, Australia

Email srecko.joksimovic@unisa.edu.au
Position Associate Professor & Co-Director, Centre for Change and Complexity in Learning (C3L)
Bio Dr. Srecko Joksimovic's research spans learning analytics, learning sciences, and responsible AI in education. He leads interdisciplinary projects with academic and industry partners, including collaborations with Southern New Hampshire University and Amazon AWS. His work explores how AI and multimodal data can enhance collaborative learning and inform educational policy. Dr. Joksimovic is active in international research networks and co-leads the EATEL SIG on Responsible and Equitable AI.

Dr. Andrew Zamecnik, University of South Australia, Australia

Email andrew.zamecnik@unisa.edu.au
Position Researcher, Centre for Change and Complexity in Learning (C3L)
Bio Dr. Andrew Zamecnik's research focuses on team dynamics—such as trust and cohesion—in collaborative learning settings. His work combines real-time analytics and AI systems to provide instructors with actionable feedback that improves group performance and student outcomes. Dr. Zamecnik has contributed to international research collaborations and partnered with non-profits to develop equitable data frameworks that support lifelong learners.

Mohammad Amin Samadi, University of California, Irvine, USA

Email masamadi@uci.edu
Position Amin Samadi specializes in AI, NLP, and learning analytics. His research explores human-AI collaboration, developing AI teammates to support teamwork and decision-making. He has published work focuses on applying NLP to model team dynamics and improve collaborative learning, and is committed to advancing responsible and effective human-AI collaboration in learning environments.

Jaeyoon Choi, University of California, Irvine, USA

Email jaeyoon.choi@uci.edu
Position PhD Student, School of Education
Bio Jaeyoon Choi's work lies at the intersection of learning analytics, human-AI interaction, and algorithmic bias. She focuses on building machine learning and natural language processing models to examine how students interact with AI, as well as addressing and mitigating algorithmic biases to promote fairness in education.

Seehee Park, University of California, Irvine, USA

Email seeheep@uci.edu
Position PhD Student, School of Education
Bio Seehee Park's research focuses on collaborative team learning and learning analytics, with a particular emphasis on human-AI collaboration in educational contexts.

Dr Valéry Psyché, TELUQ Distance Learning University, Quebec, Canada

Email valery.psyche@teluq.ca
Bio Dr Valéry Psyché is Associate Professor in the Education Department at TÉLUQ University. She holds a Ph.D. in Cognitive Computing (UQÀM). Since the start of her career, she has been involved in national and international research projects on technology-enhanced learning, ontological engineering, AI skills development, and intelligent systems for human learning.

Dr. Mohamed El Habib LARABA, Abdelhamid Mehri Constantine 2 University, Algeria

Email habib.laraba@univ-constantine2.dz
Bio Mohamed El Habib LARABA's research centers on artificial intelligence, particularly explanatory agents, qualitative reasoning, and AI applications in complex systems. He has led and contributed to national AI research projects, including one focused on modeling contextualized explanatory agents.

Dr Emmanuel G. Blanchard, Associate Professor, Le Mans University, France

Email emmanuel.blanchard@univ-lemans.fr
Bio Dr. Blanchard's research focuses on the integration of human factors—such as cognition, affect, motivation, and culture—into the design of adaptive educational technologies. In addition to his academic contributions, Dr. Blanchard brings over a decade of experience in the startup sector. A longstanding member of the AIED community for more than twenty years, he is the founder of the Culturally Aware Tutoring Systems (CATS) workshop series, whose seventh edition was held in conjunction with AIED 2024.

2 Potential Contributors

This workshop is structured as a panel-driven and discussion-focused event rather than a traditional paper submission format. Our goal is to bring together diverse perspectives from leading experts in AI, education, and inclusion to explore key challenges and opportunities at the intersection of these fields. We are inviting expert speakers from the following domains:

- Researchers specializing in AI, education, and diversity, equity, and inclusion (DEI).
- Regulatory bodies & policymakers involved in AI policy, inclusive education and AI adoption, such as representatives from UNESCO, OECD, or EU AI Act committees.
- EdTech innovators developing AI-driven collaborative learning technologies.

3 Content and Themes

As AI moves beyond being a passive tool toward an active teammate in learning environments, this workshop will explore Human-AI Teaming as a mechanism for fostering inclusive, equitable and effective collaboration. While AI research has predominantly focused on achieving human-level intelligence, less attention has been given to how AI interacts with human teams in complex problem-solving scenarios. Understanding the interplay between humans and AI in team settings will be the key to unlocking its full potential across education, healthcare, and defence.

Key Topics: To provide a comprehensive exploration of Human-AI Teaming, this workshop will cover the following key topics:

- Machines as Teammates in Collaborative Learning How AI can move beyond automation to become an active team member that enhances engagement, inclusion, and equitable participation.
- Enhancing Inclusivity in STEM with AI Teammates How AI can help underrepresented students participate in collaborative problem-solving by modulating team dynamics.
- Designing AI Teammates for Learning What attributes should an AI teammate have to positively shape engagement and group identity formation?
- Ethical Considerations in Human-AI Teaming How can we ensure AI teammates foster belonging and avoid reinforcing systemic biases?
- Maintaining Trust in AI Teammates - What strategies can be adopted to ensure AI teammate trust is calibrated with their human partners?
- Relationship Between AI and Inclusion- How can AI contribute to or hinder inclusive practices in collaborative learning environments?
- Cultural and Contextual Responsiveness - How can AI teammates enhance cultural and contextual responsiveness, ensuring inclusivity for diverse learners?

This workshop aligns with the Machines as Teammates (MaT) framework (Seeber et al., 2019), which highlights the dualities of AI as a teammate vs. tool, its role in shaping team collaboration, and the potential risks and benefits of AI-driven team participation [6].

4 Significance and Relevance to the AIED Community

The evolving narrative in AI in education has shifted from automation to augmentation, highlighting AI's role as a collaborative partner rather than a standalone tool. This workshop aligns with this transition by exploring how AI teammates can actively support and enhance human learning interactions. Instead of merely providing recommendations or automating tasks, AI-driven collaborators can foster deeper engagement, facilitate equitable participation, and contribute to more inclusive and effective learning environments.

While current AI systems primarily serve as tools for learning, advancements in Generative AI, Natural Language Processing (NLP), and adaptive learning open new possibilities for AI teammates. These systems can analyze group dynamics in real time, promoting balanced participation and mitigating dominant voices that may overshadow others. They can assist teachers and students in navigating complex collaborative interactions, providing timely guidance and insights. Furthermore, AI-driven teammates can intervene when inclusivity is at risk, ensuring that diverse perspectives are acknowledged and fostering more equitable learning experiences.

The relevance of AI teammates in education is particularly critical in STEM fields and team-based problem-solving, where disparities in representation and participation can significantly impact learning outcomes. AI teammates are not intended to replace human decision-making but rather to enhance teamwork, providing scaffolding and adaptive support where needed. By facilitating more inclusive and dynamic interactions, they hold the potential to transform learning environments into spaces where all students can actively contribute and thrive.

This workshop will bridge AI and social science research to explore how AI-enhanced teammates can be designed, deployed, and evaluated in educational contexts. By bringing together researchers, educators, policymakers, and EdTech developers, we aim to advance discussions on the role of AI as a collaborative agent and examine strategies for ensuring these technologies are both effective and ethically aligned with human-centered learning goals.

5 Workshop Format and Activities

This interactive workshop will center on panel discussions and participant-driven dialogue. The structure will include:

– Expert Presentations Concise talks from leading researchers in AI, education, and DEI, providing insights into key challenges and advancements.

- Panel Discussion A moderated conversation with experts exploring human-AI teaming, collaboration, and inclusivity in AI-enhanced learning environments.
- Interactive Discussion Sessions Small-group discussions to exchange perspectives, pose critical questions, and generate ideas for future directions.
- Live Demonstrations Showcasing AI tools designed to explore human-AI interaction in educational settings.

6 Target Audience and Expected Participation

This workshop is designed for a diverse audience engaged in the intersection of AI and education, including:

- AI and education researchers and policymakers exploring AI-driven collaborative learning and strategies for bias mitigation.
- Educators and instructional designers seeking equitable AI solutions to enhance classroom learning experiences.
- EdTech developers and AI engineers focused on designing human-centered AI tools that support inclusive learning.

We expect 2030 participants, fostering a highly interactive and engaging environment that encourages meaningful discussions and collaboration.

7 Previous Editions

While this is a new workshop at AIED, it builds on prior research and discussions in the field of AI-driven collaboration and inclusivity in education. Specifically, it draws from:

- Nixon et al. (2024) on AI-driven interventions for inclusive STEM teamwork [3].
- Seeber et al. (2019) Machines as Teammates framework, which provides foundational insights into AI collaboration research [6].
- Recent work by Joksimovic and colleagues (2023) on exploring the role AI in complex problem-solving [2].
- The workshop on Responsible and Ethical AI in Education (https://react.c3l.ai), which focuses on fairness, transparency, and accountability in AI-enhanced learning environments, delivered by Joksimovic and colleagues at ECTEL 2024.
- The ongoing research by Zamecnik and colleagues (submitted) [7] on trust in human-AI teams, examining how trust emerges, evolves, and influences collaboration in AI-enhanced learning environments.
- Samadi's ongoing work in developing a research platform that facilitates human-AI collaboration and experiments through customizable collaborative personas [5].
- Psyché, V., Savard, I., Mizoguchi, R., & Bourdeau, J. (2018): Workshop on Context & Culture awareness in AI machines at ITS 2018 [1].
- Blanchard's paper on cultural intelligence in AI machines presented at AIED 2025 [4]

8 Plans for Funding Solicitation

We will explore partnerships with organizations focused on AI in education, equity, and human-AI teaming to support speaker travel and workshop logistics. However, external funding is not required for workshop execution.

9 Publishing Workshop Outcomes

As this workshop is discussion-based rather than paper-driven, key insights and recommendations will be shared through accessible formats to maximize impact. Specifically, we plan to:

- Publish a publicly available summary outlining key discussions, findings, and recommendations on an open-access platform.
- Explore potential collaborations with BOLD (Blog on Learning and Development) and AI-focused education networks to extend the reach of workshop outcomes to a broader audience.
- Maintain a dedicated workshop website, where session materials, speaker contributions, and follow-up discussions will be made available for continued engagement.

10 Advertising Plan

To ensure broad participation, the workshop will be promoted through multiple channels, targeting researchers, practitioners, and industry professionals engaged in AI and education. Our outreach strategy includes:

- Conference and Community Networks Announcements via AIED mailing lists, official conference channels, and relevant academic forums.
- Social Media Engagement Promotion across platforms such as Bluesky, Twitter, LinkedIn, and other professional networks to reach a wider audience.
- Targeted Outreach Direct engagement with AI and education research communities, including SoLAR, EdTech developers, and organizations focused on human-AI collaboration in learning.

11 Conclusion

As AI continues to evolve in education, its role is shifting from a passive tool to an active teammate, capable of supporting and enhancing collaborative learning. This workshop provides a platform for advancing research and discussion on the potential of Human-AI teaming to foster more inclusive, equitable, and effective learning environments.

Through expert insights, interactive discussions, and hands-on strategy development, we aim to address key questions, including:

1. How can AI collaborate with students and educators in meaningful and pedagogically sound ways?
2. How can Human-AI teaming contribute to more inclusive and equitable learning experiences?
3. What are the critical design and evaluation principles for AI-driven teammates in education?
4. How can we build and maintain appropriate trust in AI teammates to ensure effective collaboration?

The insights generated will contribute to ongoing discussions in AI and education, paving the way for responsible and impactful AI integration in learning environments—where trust, inclusivity, and ethical considerations remain at the forefront.

Acknowledgments. This workshop is supported by the Jacobs Foundation, grant number 2024-1533-00.

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

References

1. Blanchard, E.G., Mohammed, P.: On cultural intelligence in LLM-based chatbots: implications for artificial intelligence in education. In: International Conference on Artificial Intelligence in Education, pp. 439–453. Springer (2024)
2. Joksimovic, S., Ifenthaler, D., Marrone, R., De Laat, M., Siemens, G.: Opportunities of artificial intelligence for supporting complex problem-solving: findings from a scoping review. Comput. Educ. Artif. Intell. **4**, 100138 (2023) https://doi.org/10.1016/j.caeai.2023.100138, https://www.sciencedirect.com/science/article/pii/S2666920X23000176
3. Nixon, N., Lin, Y., Snow, L.: Catalyzing equity in stem teams: harnessing generative AI for inclusion and diversity. Policy Insights Behav. Brain Sci. **11**(1), 85–92 (2024)
4. Psyché, V., Savard, I., Mizoguchi, R., Bourdeau, J.: Workshop preface: international workshop on context and culture in intelligent tutoring systems. Intell. Tutoring Syst. (2018)
5. Samadi, M.A., JaQuay, S., Gu, J., Nixon, N.: The AI collaborator: Bridging human-AI interaction in educational and professional settings. arXiv preprint arXiv:2405.10460 (2024)
6. Seeber, I., et al.: Machines as teammates: a research agenda on AI in team collaboration. Inf. Manage. **57**(2), 103174 (2020)
7. Zamecnik, A., Febriantoro, W., Cukurova, M., Joksimovic, S., Grossmann, G., Siemens, G.: The double-edged nature of trust in human autonomous teams: a scoping review. Comput. Hum. Behav. (2025)

Preparing Teachers for an Equitable AI Integration in Education

Victoria Abou-Khalil[1(✉)], Lena Hollenstein[2], Valentina Ferraioli[1], Marius Vogt[2], Brendan Flanagan[3], Franziska Vogt[2], and Francesco Mondada[1]

[1] Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, VD, Switzerland
victoria.aboukhalil@epfl.ch
[2] St. Gallen University of Teacher Education, Saint Gallen, Switzerland
[3] Graduate School of Informatics, Kyoto University, Kyoto, Japan

Abstract. This workshop proposal aims to inform the design of teacher training on equity in the use and teaching of AI. AI is transforming both teaching practices and student learning environments and learning content. Teachers are increasingly using AI tools for course planning, some AI-based tools are being integrated into schools (e.g., for language learning), and educators are also responsible for teaching AI concepts. While equity has been recognized as a critical concern, potential equity issues are complex and depend on various factors, including school level, student demographics, the type of AI used (generative or predictive), and whether AI tools support student learning or assist teachers in planning and assessment. Additionally, inequities may emerge when teaching AI. A comprehensive understanding of equity issues is essential to inform teacher training that provides teachers with equity issues that emerge in their specific context and equips them with the necessary skills to use AI responsibly in the classroom. This workshop aims to develop such an overview by bringing together a multidisciplinary group of researchers and practitioners with experience with AI education, professional development, AI fairness, predictive and generative AI biases, and inequities in digital education. Through case studies and discussions, the workshop will identify and categorize various equity challenges, laying the foundation for future research on AI teacher training and informing the development of context-specific guidelines for equity for AIED.

Keywords: Equity · Teacher Training · AI fairness · AI Education · AIED

1 Introduction

Teachers are confronted with AI in multiple ways: some of them must teach AI concepts to students as part of digital literacy and computational thinking curricula, others must navigate students' use of AI-based tools across disciplines such as language learning and STEM education, and they are increasingly relying on AI-driven applications for assessment, personalized instruction [1] and course preparation [4]. These shifts in teaching practices create both opportunities and challenges, particularly regarding equity. While AI contains the promise of supporting students' diverse needs and bridging the divide,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 249–254, 2025.
https://doi.org/10.1007/978-3-031-99267-4_31

its integration into educational settings also raises significant equity concerns, particularly regarding access to AI tools and resources, data biases, algorithmic fairness, and the interplay between AI decision-making and human decision making as the ways AI tools present information can either challenge or reinforce teachers' existing beliefs [5]. Additionally, biases can also appear when teaching AI to children as there is a high risk of unequal learning opportunities for different groups of students [9].

While policymaking and regulations play a crucial role in shaping equitable AI adoption, educators also need to have the skills and tools to critically assess AI tools and ensure they are used in ways that promote fairness and inclusivity [7]. The UNESCO AI Competency Framework for Teachers highlights equity as a key competency, emphasizing that educators must be equipped to ensure fair AI integration in their teaching practices and decision-making processes [2]. The framework underscores the need for teachers to develop a critical understanding of AI's potential impact on learning, including issues related to bias, inclusivity, and accessibility. Training programs focused on AI literacy, ethical considerations, and equity-driven pedagogical strategies can help teachers integrate AI effectively while mitigating its risks.

To design teacher trainings that are relevant to teachers, potential equity issues should be mapped out. Equity is a context-specific issue, shaped by factors such as school level, subject discipline, specific use case scenarios, and the demographics of the student population [3, 6, 8]. Understanding the nuances of equity concerns in different contexts is crucial for developing effective solutions. Despite its importance, discussions regarding equity issues of AI integration and teaching have been contextualized and have been focused on generalized possible risks and challenges. Creating an overview of equity issues that could be used to inform teachers and create guidelines is a complex task that requires pedagogical, technical, and domain knowledge as well as the perspectives of institutions and decision-makers.

2 Workshop Objectives

This workshop aims to bring together a multidisciplinary group of researchers, practitioners, policymakers, and EdTech developers to map out a comprehensive overview of these varying equity challenges and the contextual; factors that should be considered. By exploring case studies and engaging in discussions, participants will work toward mapping the key equity issues surrounding AI adoption, including differences across school levels, student demographics, and types of AI applications used in teaching and learning. The workshop will provide a foundation for future research and contribute to the development of equity-driven AI teacher training programs, ultimately supporting more just and effective AI integration in schools.

3 Expected Outcomes

We envision that this workshop will result in three main outcomes. First, through the discussions and case studies, we will create an overview of the different equity issues that should be integrated into teacher training for the different areas: using AI for teaching, using AI to support student learning, and teaching AI concepts. We will analyze

each of the areas through the lens of access, data biases, algorithmic fairness, and the interplay between AI decision-making and human decision. Second, we will use the multidisciplinary perspectives of our participants to create an overview of the contextual factors that should be considered when creating content regarding equity for teacher training (e.g., education level, subject area, student demographics, etc.). The first two objectives will contribute to the creation of a framework for integrating equity concepts into teacher training for AIED. The third outcome is to create a collaborative research agenda that identifies future research direction and collaborate on a common publication that illustrates the workshop discussions and results.

4 Relevance and Importance to the AIED Community

This workshop is important for the AIED community as it addresses the major challenge of setting the ground for equity-centered teacher training. Aligning with AIED 2025's theme of Human-Centered AI in Education, this workshop ensures that teachers are aware of equity considerations when using AI tools or teaching AI and are prepared to critically assess and use AI to support fair and inclusive AI education. This workshop also aims to bridge the gap between AIED researchers and teacher training specialists, a community that has been somewhat distant from AIED but is essential for the effective and just integration of AI in classrooms.

By bringing together researchers, practitioners, and policymakers, this workshop will map key equity challenges in AI adoption across different school contexts, laying the foundation for future research and AI teacher training frameworks. This is a critical step toward ensuring AI-driven education that supports teachers with autonomy in decision-making while giving them the tools for a more equitable use and teaching of AI.

5 Content and Themes

The call for participation will welcome theoretical or empirical submissions of position papers, case studies, or ongoing research on the following topics in the K12 context. Submissions focusing on the primary school context are particularly welcome.

- Algorithmic and data biases in AI systems and their impact on equity in education
- Algorithmic fairness in AI systems used in teaching and learning environments
- The interplay between AI decision-making and human decision-making and its impact on fairness
- Equity issues and challenges when teaching AI concepts to students
- Teacher competency frameworks focused on equity in AI integration and usage
- Domain-specific AI-related equity issues (e.g., in subjects like mathematics, writing, etc.)
- Teachers' and students' views on potential equity issues in AI education and its integration into classrooms
- Policymaking for equity in AI use in education: developing frameworks to ensure fair and inclusive AI adoption

6 Expected Target Audience

We expect participants to be multidisciplinary and come from diverse fields such as learning analytics, instructional design, teacher training, equity, and AI. We also particularly welcome EdTech developers, K12 teachers, and other practitioners as well as decision-makers. The multidisciplinary of participants is essential to achieve the objectives of the workshops and and will make sure to target different groups through the organizers' diverse networks. We plan to have around 20 participants be part of the workshop, excluding the workshop organizers. Participation in the workshop will be open, which means that attendees don't need to submit a paper and make a presentation. The workshop organizers will select presenters based on the quality of submissions, to ensure a well-rounded selection of topics and a balance between research papers, practitioner papers, and position papers.

The organizing team is multidisciplinary and includes experts from the fields of AI curriculum, teacher digital professional development, and educational technologies. We will distribute the call for participation in our respective networks and publicize it through a website dedicated to the workshop.

7 Format and Activities

We propose a half-day workshop with a panel discussion to present equity considerations from different perspectives, paper presentations, activities, and discussions. We will also support online participants. Table 1 provides a detailed tentative schedule.

The Panel Discussion: The panel will bring together experts to share their perspectives on equity considerations in AIED. It will feature three speakers from the fields of AI in education, AI education, and digital teacher training. Potential speakers include Professor Tanja Kaeser, specializing in educational data mining and AI in education at EPFL, and Professor Kenneth Horvath, an expert in educational sciences with a focus on digital educational inequalities at PH Zürich.

The Presentations: The format of the presentations will be in the style of lightning talks, with presenters allotted 5 min to present their papers. The presentations will be organized into groups based on similar themes, and after each presentation, there will be a 30-min period for questions and discussions. The organizers will provide guiding questions to facilitate the discussions.

Group Activity: During the workshop, the participants will be organized into smaller groups of 4 to 5 individuals. Each group will be provided with a specific topic related to a particular context such as personalization, teaching AI, and using AI-supported tools. Each group will be presented with case studies of real-world situations and applications that involve AI and are currently in use in schools. The groups will then be asked to brainstorm potential equity issues and contextual factors that should be considered in each case. The brainstorming results should be reported in a table. Online participants will form a group via Zoom to complete the activity. Following the construction of the overview we will dedicate time to the sharing of the groups' work and the discussion of their results. The group activity will allow participants to engage in a multidisciplinary

discussion of the various challenges and opportunities related to equity in the context of AIED. If they desire, the contact between the different working groups will be supported for the potential creation of a working paper together.

8 Workshop Chairs

Victoria Abou-Khalil is a Postdoctoral Researcher at EPFL and the Executive Director of the EPFL – ETH Zurich Doctoral Program in the Learning Sciences. She also coordinates the DEEP consortium on Digital Education for Equity in Primary Schools. Her research focuses on the design and evaluation of digital education and technologies that support diverse learners, including migrants and students in low-resource environments.

Valentina Ferraioli is a PhD candidate in the EPFL – ETH Zurich Doctoral Program in the Learning Sciences. Her research centers on supporting primary school teachers in understanding and teaching core AI concepts within digital education. She explores the connections between AI and computational thinking to extend existing curricula with AI-related topics.

Marius Vogt is a Doctoral researcher at St. Gallen University of Teacher Education. He works in a project on teachers' continuous professional development on AI and equitable digital education for children in primary school. His current focus lies in developing playful approaches to foster AI literacy among children.

Brendan Flanagan is an Associate Professor at the Center for Innovative Research and Education in Data Science in the Institute for Liberal Arts and Sciences, and also holds appointments at both the Data Science Course and Social Informatics Course within the Graduate School of Informatics at Kyoto University. His research is in Learning Analytics and AI in education in both K-12 and Higher Education, with a particular focus on AI/machine learning and statistical models of reading behavior. Other areas of interest include NLP/Text Mining, Computer Assisted Language Learning, CSCL and the Application of Blockchain in Education.

Lena Hollenstein is a Senior Researcher at St.Gallen University of Teacher Education and Co-Head of Studies Master Early Childhood Studies. Her research focus is on research in early childhood education and teacher education in digital education, teacher expectations and teacher's professional competence. Her current project focuses on teachers' continuous professional development on AI and equitable digital education for children in primary school. She has worked on gender-sensitive pretend play ideas for digital transformation in kindergartens.

Franziska Vogt is the director of the Institute for Early Childhood Education 0 to 8 years at the St.Gallen University of Teacher Education. She is also responsible for the Masters Program Early Childhood Studies. Her research interests focus on teaching and learning in early childhood education settings (childcare centers, playgroup, family centers), kindergarten and primary school, with a particular focus on play, interaction quality and multiprofessional cooperation. Current projects include play and digital education and equitable education regarding gender and multilingual learners.

Francesco Mondada is Professor of Robotics at the Ecole Polytechnique Fédérale de Lausanne (EPFL) and academic director of the LEARN center. Francesco Mondada is responsible for the design of numerous mobile robots for education, marketed and used

in thousands of schools and universities. He has received several awards, including the EPFL Best Teacher Award, and the Polytechnik-Preis for the Thymio robot, recognized as the best educational project in Europe in the field of digital tools.

9 Potential Committee Members

Jade Maï Cock, EPFL, Switzerland.
Richard Lee Davis, KTH Royal Institute of Technology, Sweden.
Jessica Dehler Zufferey, EPFL, Switzerland.
Jakub Macina, ETH Zürich, Switzerland.
Rwitajit Majumdar, Kumamoto University, Japan.
Konstantinos Michos, EPFL, Switzerland.

References

1. Abou-Khalil, V., Flanagan, B., Ogata, H.: Personal vocabulary recommendation to support real life needs. In: International Conference on Artificial Intelligence in Education, pp. 18–23. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78270-2_3
2. Cukurova, M., Miao, F.: AI competency framework for teachers. UNESCO Publishing (2024)
3. Ghosh, A., Genuit, L., Reagan, M.: Characterizing intersectional group fairness with worst-case comparisons. In: Artificial Intelligence Diversity, Belonging, Equity, and Inclusion. PMLR, pp. 22–34 (2021)
4. Guo, Y., Yu, H.: Exploration of education transformation and teacher literacy in the age of artificial intelligence. In: 2023 5th International Workshop on Artificial Intelligence and Education (WAIE). IEEE, pp. 38–42 (2023)
5. Holstein, K., Doroudi, S.: Equity and artificial intelligence in education. In: The ethics of artificial intelligence in education. Routledge, pp. 151–173 (2022)
6. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: International Conference on Machine Learning. PMLR, pp. 2564–2572 (2018)
7. Pedro, F., Subosa, M., Rivas, A., Valverde, P.: Artificial intelligence in education: challenges and opportunities for sustainable development (2019)
8. Viberg, O., Kizilcec, R.F., Wise, A.F., Jivet, I., Nixon, N.: Advancing equity and inclusion in educational practices with AI -powered educational decision support systems (AI - EDSS). Brit. J. Educational. Tech. **55**, 1974–1981 (2024). https://doi.org/10.1111/bjet.13507
9. Wammes, D., Slof, B., Schot, W., Kester, L.: Teacher judgement accuracy of technical abilities in primary education. Int. J. Technol. Des. Educ. **33**, 415–438 (2023). https://doi.org/10.1007/s10798-022-09734-5

AI Literacy For All: 1st International Workshop on AI Literacy Education For All

Ruiwei Xiao[1(✉)], Ying-Jui Tseng[1], Hanqi Li[2], Guanze Liao[3], John Stamper[1], and Kenneth R. Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA
ruiweix@cs.cmu.edu
[2] New York University, New York, NY 10012, USA
[3] National Tsinghua University, Hsinchu 30013, Taiwan

Abstract. As artificial intelligence (AI) becomes increasingly embedded across societal and professional domains, AI literacy has emerged as a global educational priority. National and international policy initiatives underscore a growing consensus through executive orders and national plans equipping all individuals with the knowledge and critical capacity to engage with AI is essential for fostering equitable, ethical, and future-ready societies. Therefore, we propose the first International Workshop on AI Literacy Education for All (ALIT4ALL), aims to advance research and practice on AI literacy education for non-technical learners, including K12 students, educators across disciplines, and workforce professionals. By convening researchers, practitioners, policymakers, and industry stakeholders, the workshop seeks to address persistent challenges such as limited teacher preparation, inaccessible learning resources, fragmented upskilling programs, and the absence of validated assessment frameworks. Key topics include the design and evaluation of AI literacy curricula, professional development models for educators, inclusive pedagogical strategies, and innovative assessment methodologies. By fostering interdisciplinary collaboration, this workshop aims to generate actionable insights and strategies for equipping all learners with the necessary knowledge and skills to navigate an AI-driven world and contributing to the global effort to democratize AI understanding and participation across diverse educational and professional contexts.

Keywords: AI Literacy · Digital Literacy · Teaching AI · Curriculum Design · AI in Education · Professional Development

1 Motivation and Theme

As AI continues to reshape work and society, its growing presence has galvanized global efforts to embed AI literacy into education systems. From UNESCO's international guidance [12,13] to the U.S. *Executive Order on AI education* [10],

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 255–260, 2025.
https://doi.org/10.1007/978-3-031-99267-4_32

the EU's *Digital Education Action Plan* [2], and China's *Education Modernization Plan 20242035* [8], national and global actors are recognizing AI literacy as essential for future readiness. These efforts aim to ensure that individuals across professions and roles develop the AI literacy needed to understand both the transformative benefits and the societal risks AI poses to different stakeholders. Today's workforce, making collaborative decisions with AI, and children, who will become tomorrow's leaders, policymakers, innovators, and users, must be equipped with AI literacy to help build a responsible, informed society with sufficient understanding of AI's capabilities and limitations, ethical considerations, and societal implications [7,11]. Without purposeful education, non-technical audiences—including K12 students, teachers in non-STEM fields, and workforce professionals—may lack the foundational awareness needed to make informed decisions, exercise critical judgment, and utilize AI for personal or societal benefit.

Despite the recognized importance of AI literacy for non-technical learners, bridging the AI gap remains challenging across multiple fronts. AI concepts, often cloaked in complex jargon and grounded in advanced mathematics, can be intimidating to newcomers [16], leading to limited enrollment in training programs or superficial engagement with existing resources [5,15]. Many teachers, especially those in non-STEM fields, lack accessible professional development opportunities or clear curricular support [4], leaving them uncertain about how to introduce AI fundamentals and ethical considerations into the classroom. In the workplace, professionals of varying skill levels encounter fragmented upskilling initiatives that are not always aligned with industry needs or ethical guidelines [1,9], resulting in an uneven ability to adopt, manage, or critically evaluate AI-based tools. Furthermore, there is no widespread consensus on how to assess AI literacy, making it difficult to benchmark progress, refine instructional approaches, or ensure that AI education reaches all learners equitably [6]. These overlapping gaps underscore the need for intentional, inclusive, and ethically grounded AI literacy programs that serve diverse non-technical learners.

This workshop centers on AI Literacy Education for Non-Technical Learners, including:

K12 Students: Early exposure to AI fundamentals can spark lifelong interest, nurture critical thinking, and empower the next generation to participate fully in an AI-driven society [14].

Teachers: As multipliers of knowledge, teachers need the professional development and resources to confidently integrate AI concepts into the classroom [4]—regardless of subject matter—while also guiding students to think critically and ethically about AI's impact.

Workforce Training: In an evolving job market, professionals at various skill levels need continuous upskilling to navigate AI tools responsibly and effec-

tively [3], addressing ethical concerns such as bias, data privacy, and algorithmic accountability.

Through collaboration among educators, researchers, policymakers, and industry partners, we can co-create scalable, innovative, and ethically grounded AI literacy programs that empower every learner—regardless of background—to navigate and shape the AI-powered future.

2 Objectives

This workshop aims to bring together researchers, educators, industry professionals, and policymakers to explore comprehensive strategies for AI Literacy education that serve non-technical learners: including students and teachers in K-12 and higher education, and the broader workforce. The objectives of the workshop are to:

- Design and Evaluation of AI Literacy Programs
- Educational Technologies for AI Literacy Development
- AI Literacy Assessment
- AI Literacy Pedagogies for Non-Technical Learners
- Diversity, Equity, and Inclusion in AI Literacy

3 Call for Submission

While submission is not required to attend, we strongly encourage scholars, educators, practitioners, and industry professionals to contribute 4-6 page work-in-progress or position papers related to AI Literacy Education for Non-Technical Learners. Submissions may include research on K–12 curricula, teacher training programs, workforce upskilling strategies, ethical frameworks, policy considerations, or novel assessment and educational technology approaches.

The Call for Papers opens on March 28, 2025, and the submission deadline is May 19, 2025. Each manuscript will undergo a double-blind review by at least two program committee members, and authors will be notified of acceptance decisions on June 2, 2025. Accepted papers will be highlighted in workshop sessions designed to share state-of-the-art practices, methodologies, and forward-looking perspectives that help close the AI literacy gap for non-technical audiences.

We particularly invite submissions that:

- Investigate frameworks and methodologies for implementing AI literacy, ensuring broad accessibility and meaningful impact across diverse learning contexts.
- Highlight and evaluate emerging tools and platforms that enable hands-on AI experiences and conceptual understanding for non-technical audiences.
- Advance new paradigms for measuring AI literacy, including project-based tasks, reflective portfolios, and performance-based evaluations that capture both practical skills and ethical considerations.

- Explore inclusive teaching approaches that demystify core AI concepts for K–12 students, educators, and professionals seeking upskilling, regardless of subject-area expertise.
- Identify and address systemic and cultural barriers limiting access to AI education, proposing strategies and policies to foster equitable participation and representation in AI-driven fields.

We welcome papers presenting empirical research, theoretical discussions, design prototypes, or reflections on work-in-progress, and encourage diverse methodological approaches (qualitative, quantitative, design-based, or mixed methods). Submissions must follow the workshop's paper template; accepted papers will be published in the workshop proceedings.

By gathering a wide range of perspectives – from researchers and educational technologists to policymakers, teachers, and industry stakeholders – this workshop aims to generate actionable insights that make AI literacy more inclusive, engaging, and impactful. We look forward to your contributions as we work together to shape the future of AI Literacy Education for non-technical learners in an equitable and ethically grounded manner.

4 Workshop Format and Activities

This workshop will be a half-day interactive event designed to engage participants in hands-on exploration, discussion, and collaboration around AI Literacy Education for Non-Technical Learners. The provisional schedule is outlined below:

- **Introductions**: Workshop organizers and participants will briefly introduce themselves, followed by a concise overview of the workshop's focus on bridging the AI literacy gap among diverse, non-technical audiences.
- **Invited Guest Speaker**: A leading expert in AI literacy (not on the organizing committee) will deliver a keynote on pivotal challenges and emerging best practices for designing inclusive AI education across K–12 classrooms, teacher training, and workforce development.
- **Short Presentations**: Authors of accepted submissions will present their work in brief, focused segments, followed by Q&A. Presentations may address any workshop-related topic, including scalable AI literacy programs, assessment innovations, or case studies highlighting inclusive pedagogies.
- **Round-Table Discussions**: Attendees will rotate among topic-specific tables (for example, K–12, teacher training, workforce upskilling, ethics and inclusion, emerging technologies), reflecting the workshop's objectives. Each discussion will emphasize practical solutions, challenges encountered, and opportunities for collaboration in advancing AI literacy initiatives.
- **Mini-Hackathon**: Participants will collaborate in small, interdisciplinary teams to prototype or brainstorm new tools, curricular frameworks, policy proposals, or resources aimed at fostering AI literacy among non-technical learners, encouraging creative thinking about accessibility, engagement, and ethical considerations.

- **Concluding Remarks and Community Engagement**: Workshop organizers will synthesize key insights from the day's activities and discuss potential next steps for collaboration, resource sharing, or research partnerships. Future engagement will be facilitated via a dedicated Slack channel and mailing list.

All accepted workshop papers will be published in a CEUR proceedings volume. To promote ongoing engagement, the workshop website will be updated with key takeaways, links to relevant materials, and participant contact information, ensuring a lasting community of practice around AI Literacy Education.

5 Organization

- **Ruiwei Xiao** is a PhD student at Carnegie Mellon University (CMU) and is advised by Dr. John Stamper, and co-founder of Active AI. Her research and startup experience are focused on AI Literacy and Computing Education.
- **Ying-Jui Tseng** is a CMU alumni and a co-founder of Active AI. With years of UX design and learning design experiences for companies such as Amazon, Coursera and PaGamO. He founded Active AI and has delivered AI literacy solutions that have benefited thousands of learners in Taiwan.
- **Hanqi Li** is an incoming PhD student at Department of Education, UCSD. She will be advised by Amy Eguchi, who takes part in the AI for K-12 initiative, jointly sponsored by AAAI and CSTA, as an advisory group member, working collaboratively with K-12 classroom teachers teaching CS who also took part in the development of the 2017 CSTA standards.
- **Guanze Liao** is the Dean of the National Tsinghua University School of Education in Hsinchu, Taiwan. His work focuses on AI literacy education in teacher training programs and integrating AI into educational contexts. Additionally, his leadership and strong network within Taiwan's teaching community play a crucial role in advancing AI education.
- **John Stamper** is an Associate Professor at the Human-Computer Interaction Institute at Carnegie Mellon University and the Technical Director of the Pittsburgh Science of Learning Center DataShop. His work involves leveraging educational data mining techniques and the creation of data tools.
- **Ken Koedinger** is a Professor of Human-Computer Interaction and Psychology at Carnegie Mellon University. His research has contributed to an understanding of student thinking and learning that has implications for educational technology and teaching. He has created cognitive models that have been used to create educational materials and technologies, including intelligent tutoring systems that adapt to student needs. His work combines cognitive science, artificial intelligence, and educational practice.

References

1. Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., Knight, S.: Explicating AI literacy of employees at digital workplaces. IEEE Trans. Eng. Manage. **71**, 810–823 (2022)
2. European Commission: Digital education action plan (2021–2027). https://education.ec.europa.eu/focus-topics/digital-education/action-plan (2021), policy Framework
3. Kaspersen, M.H., et al.: From primary education to premium workforce: Drawing on k-12 approaches for developing AI literacy. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2024)
4. Lee, I., et al.: AI book club: an innovative professional development model for AI education. In: Proceedings of the 53rd ACM Technical Symposium on Computer Science Education, vol. 1, pp. 202–208 (2022)
5. 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)
6. Lintner, T.: A systematic review of AI literacy scales. npj Sci. Learn. **9**(1), 50 (2024)
7. 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)
8. Ministry of Education of the People's Republic of China: Education modernization plan 2024–2035 (2025). https://www.gov.cn/zhengce/202501/content_7000579.htm
9. Southworth, J., et al.: Developing a model for AI across the curriculum: transforming the higher education landscape via innovation in AI literacy. Comput. Educ. Artif. Intell. **4**, 100127 (2023)
10. The White House: Advancing artificial intelligence education for American youth (2025). https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth/
11. 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)
12. UNESCO: Artificial intelligence and education: Guidance for policy-makers (2021). https://unesdoc.unesco.org/ark:/48223/pf0000376709
13. UNESCO: Guidance for generative AI in education and research (2023). https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research
14. Williams, R., et al.: AI+ ethics curricula for middle school youth: lessons learned from three project-based curricula. Int. J. Artif. Intell. Educ. **33**(2), 325–383 (2023)
15. Xia, Q., Chiu, T.K., Chai, C.S.: The moderating effects of gender and need satisfaction on self-regulated learning through artificial intelligence (AI). Educ. Inf. Technol. **28**(7), 8691–8713 (2023)
16. Yang, W.: Artificial intelligence education for young children: why, what, and how in curriculum design and implementation. Comput. Educ. Artif. Intell. **3**, 100061 (2022)



Workshop on Advancements in AI-Supported Exploratory Learning (AI4EL)

Manolis Mavrikis[1], Sébastien Lallé[2(✉)], Roger Azevedo[3], Gautam Biswas[4], Jade Mai Cock[5], Cristina Conati[6], Sokratis Karkalas[7], Tanja Kaser[5], James Lester[8], Ido Roll[9], Janan Saba[10], and Megan Wiedbusch[3]

[1] UCL Knowledge Lab, University College London, London, UK
m.mavrikis@ucl.ac.uk
[2] Sorbonne Université, CNRS, LIP6, 75005 Paris, France
sebastien.lalle@lip6.fr
[3] School of Modeling Simulation and Training, University of Central Florida, Orlando, USA
[4] Department of Computer Science, Vanderbilt University, Nashville, TN, USA
[5] Machine Learning for Education, EPFL, Lausanne, Switzerland
jade.cock@epfl.ch
[6] Department of Computer Science, The University of British Columbia, Vancouver, BC, Canada
[7] Computing and Mathematics, University of Derby, Derby, Derbyshire, UK
[8] NC State University, Raleigh, NC, USA
[9] Faculty of Education in Science and Technology, Technion Israel Institute of Technology, Haifa, Israel
[10] School of Education, Hebrew, University of Jerusalem, Jerusalem, Israel
janan.saba@mail.huji.ac.il

Abstract. This workshop aims to provide a venue for researchers to discuss latest work on AI-based support for Exploratory Learning Environments (ELE). It builds on a series of workshops in this area and latest advancements from the organising committee but also several colleagues who have indicated an interest in this area.

Keywords: exploratory learning · inquiry learning · vocational learning · open-ended learning · metacognition · learner modelling · intelligent scaffolding

1 Introduction

About a decade ago a series of workshops on Intelligent Support for Exploratory and Open-Ended Learning Environments in several AIED, EC-TEL and other conferences were a significant platform within the community, to share research on the challenges and opportunities of AI and analytics in supporting learners and educators within what we broadly refer to as 'exploratory learning'

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 261–266, 2025.
https://doi.org/10.1007/978-3-031-99267-4_33

(e.g., [20])[1]. The term helped us capture engagement with open-ended, interactive digital environments that adhere to constructionist perspectives, inquiry activities, vocational learning and other similar approaches. These environments typically include learning simulations, microworlds, scientific inquiry tools, and open-world educational games. Because these environments allow learners to engage with learning in flexible and self-directed ways while exploring and solving complex problems, they often foster higher-order cognitive, metacognitive, and emotion regulation skills [20,21].

This proposed workshop builds on that foundational work in light of recent advancements in pedagogy, instructional design on one hand, and adaptive support, multimodal analytics, and generative and other AI approaches on the other hand. Notably, AI approaches can serve multiple roles in exploratory learning environments (ELEs), e.g., to model learners' cognitive, metacognitive, motivational, and affective processes; to generate personalised activities; to offer adaptive support; to interpret rich text and data for research purposes.

In alignment with the AIED 2025 theme, the workshop will primarily examine how AI can support learning and teaching in this context while ensuring it is designed equitably, responsibly, maintaining the essence of empowerment that exploratory learning aims to foster for teachers and students. The organisers have recently advanced the field in several areas e.g., equitable design [7], mitigating bias in behavioural models [4], machine learning (including generative AI and LLMs) [5,9,11], knowledge representation [10], explainability [6,19], digital twins [14], and cognitive and metacognitive scaffolding [2,3,15]. Discussions with colleagues in the field (see a recent review [20]) indicate a need for a venue to report progress in light of recent advances and spark collaborations. This workshop in AIED therefore is timely and acts as a forum to discuss these topics and identify future directions.

2 Themes

We identified eight key themes for the workshop, presented below.

- **Pedagogical Strategies in ELEs.** The strongest design intention of ELE is the playfulness and freedom with which they provide learners, so that they can investigate the system while engaging with it as a personally meaningful activity. However, it has been observed that without the implementation of appropriate pedagogical strategies, ELEs may produce suboptimal learning outcomes, often resulting in students experiencing difficulty in interpreting key concepts during unguided discovery processes. Therefore, the incorporation of specific pedagogical strategies is essential to achieve a balance between free exploration and effective learning. For instance, empirical studies indicate that the Problem-Solving Prior to Instruction (PS-I) approach can be particularly effective. This

[1] See for example M. Mavrikis et al., Intelligent Support in Exploratory and Open-ended Learning Environments in AIED 2015. https://ceur-ws.org/Vol-1432/iseole_laels_proc.pdf.

strategy involves engaging learners in exploratory learning within ELEs, followed by direct instruction on the target subject matter. Such an approach has demonstrated promise in overcoming the aforementioned challenges, particularly in fostering conceptual understanding and facilitating knowledge transfer [13]. The guided-learning and inquiry-learning literature offers important insights for automating support without undermining learner agency.

- **User model, domain model... activity model?** Given the aforementioned challenges, a classical approach to learner and domain modelling based on concepts does not always fit the purposes of ELEs. The classic scenario, in which learners are required to study materials about a concept and then their knowledge level is assessed through testing, is not appropriate in many cases so the modelling an open-ended activity is a challenging problem. There is also no clear set of strategies for solving an exploratory task; different approaches can all be valid, and it is challenging to distinguish a correct one from an incorrect one, except for trivial cases.

- **(Multimodal) data analysis.** In order to understand the actions of the learner, machine learning and data mining techniques can be used to analyse the data provided by the system. This data can be multimodal, integrating sources such as eye tracking, speech, physiological sensors, and interaction logs, to offer a richer and more nuanced picture of learner behaviour [1,17]. Applying these techniques in an open environment, where the number of possible courses of actions is much higher and the data is increasingly heterogeneous and multimodal, poses additional challenges.

- **Explainability, transparency and trustworthiness.** Explainability is inherently challenging in this context as there are no a-priori defined skills and strategies and there are many possible courses of actions, making the definition of traditional models challenging. Clickstream-based as well as sensor-based models constitute a promising solution, but they are not interpretable by design. Explainable AI can help here. However, existing post-hoc explainers tend to disagree [18] and in-hoc interpretation of neural networks models tends to be low-level [5]. While more traditional approaches have allowed insights into student behaviours in these environments [16] and in authoring and communicating the design to stakeholders [10], there is a compelling need for designing inherently interpretable models. Another avenue of research is in examining whether and how to personalize these explanations to the different stakeholders and user needs [6]. Additionally, human-in-the-loop approaches may be offering promising directions, but their applicability in this context remains underexplored.

- **Authoring and teacher-designed systems.** How can AI-supported authoring tools enhance educators' understanding of student progress while also improving the transparency of learning processes? When explainability is framed in the context of rule-based AI or other teacher-authored systems, it shifts from the traditional focus on justifying AI outputs to something deeper. Essentially, explainability in this context is not just about transparency for students, but also about making the AI-driven learning process more intelligible and

reflective for educators. It allows them to refine both their pedagogical strategies and the interactive systems through which students learn. [12]

- **Equity and personalisation.** In highly interactive, open-ended learning environments, student trajectories are inherently non-deterministic. Learner activity indicators have the potential to provide rich, nuanced insights into student development throughout the learning process. In this context, how can personalisation be leveraged to address equity? What challenges and opportunities arise when tailoring learning experiences to diverse student needs while ensuring fairness and inclusivity?

- **Teachers need support too.** It is unrealistic to assume that a classroom teacher will be able to attend to the needs of all their students. Their work would therefore become easier if they were supported during the use of such environments in classroom. Therefore, intelligent support and information for the teacher (e.g. visualisation of students' activities) become paramount.

- **The role of generative AI (GenAI).** The emergence of GenAI tools, in particular LLMs and conversational chatbots, presents exciting opportunities for advancing research on ELEs [8,9,19]. In this context, these tools can serve not as a replacement for student thinking but as a guide, co-learner or co-designer in the learning process. However, integrating Generative AI tools into ELEs presents new challenges, such as ensuring pedagogically sound interactions, mitigating hallucinations, balancing AI assistance with productive struggle, and enhancing chatbots' ability to perceive and interpret the ELE's dynamic content and context. How can we rethink assessment and pedagogy to leverage GenAI for deeper exploratory learning rather than simple automation of tasks?

3 Format of the Workshop

Following the success of previous workshops, we proposed to use a question-oriented organisation with lively discussions rather than a mini-conference. The methodology used is based on an adaptation of the 'learning discussion forum' (aka 'Learning Cafe Methodology') which involves experts' introductions and group discussions ensuring that all participants can have a direct impact in addressing the workshop questions.

The working methodology for the workshop is made up of the following steps:

- A short presentation is made that raises some open problems in the field and post challenging questions that the participants of the workshop have to answer in advance. Those interested in the workshop submit their papers, covering some of the challenging questions rose by the experts. Each paper is reviewed by members of the Program Committee.
- As Coordinators we will set up a collaborative environment such as Slack or Discord to facilitate open discussion among workshop participants previous to the workshop day. The authors of the papers send the final version of their paper, which is shared with other contributors
- Open discussion among participants takes place in advance of the workshop day engaging also colleagues who cannot make it in the workshop.

On the workshop day, depending on the topics covered in the papers and the number of participants, two or three round tables are settled, each managed by a Moderator. The tables have A-1 sheets and markers. A brief overview of each of the topics is done, raising the challenging questions. Each table brainstorms led by the Moderator who at the end presents the conclusions to the audience. The Coordinators summarize the conclusions with the collaborative help of the group. Moreover, the community of interest set up in the collaborative platform for the workshop will be kept and extended with people interested in the topics. After the roundtables, an interactive demo session takes place, where participants can showcase innovative technologies in ELEs. Final conclusions are structured and uploaded in the website[2] to be shared with the AIED community and, depending on interest, we will consider a special issue.

Acknowledgments. This research is co-funded by the European Union HORIZON-WIDERA-2021-ACCESS-03-01 (Grant no: 101078875) and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe Guarantee funding scheme.

References

1. Ashwin, T.S., Snyder, C., Akpanoko, C.E., Srigowri, M.P., Biswas, G.: Combining multimodal analyses of students' emotional and cognitive states to understand their learning behaviors. In: International Conference on Computers in Education (2024). https://library.apsce.net/index.php/ICCE/article/view/4854
2. Azevedo, R., et al.: A taxonomy for enhancing metacognitive adaptivity and personalization in serious games using multimodal trace data. In: Joint International Conference on Serious Games, vol. 15259, pp. 27–40. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-74138-8_3
3. Branthôme, M., Lallé, S.: Impact of adaptive feedback on learning programming with a serious game in high schools' classes. In: 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2025) (2025). https://doi.org/10.1145/3699682.3728326
4. Cock, J.M., Bilal, M., Davis, R., Marras, M., Kaser, T.: Protected attributes tell us who, behavior tells us how: a comparison of demographic and behavioral oversampling for fair student success modeling. In: 13th International Learning Analytics and Knowledge Conference, pp. 488–498. ACM (2023). https://doi.org/10.1145/3576050.3576149
5. Cock, J.M., Marras, M., Giang, C., Käser, T.: Generalisable methods for early prediction in interactive simulations for education. arXiv:2207.01457 (2022)
6. Conati, C., Barral, O., Putnam, V., Rieger, L.: Toward personalized XAI: a case study in intelligent tutoring systems. Artif. Intell. **298**, 103503 (2021). https://doi.org/10.1016/j.artint.2021.103503
7. Fonteles, J.H., Akpanoko, C.E., Wisniewski, P.J., Biswas, G.: Promoting equitable learning outcomes for underserved students in open-ended learning environments. In: Proceedings of the 23rd Annual ACM Interaction Design and Children Conference, pp. 307–321. ACM (2024). https://doi.org/10.1145/3628516.3655753

[2] https://transeet.eu/ai4exl/.

8. Giannakos, M., et al.: The promise and challenges of generative AI in education. Behav. Inf. Technol. 1–27 (2024). https://doi.org/10.1080/0144929X.2024.2394886
9. Goslen, A., Kim, Y.J., Rowe, J., Lester, J.: LLM-based student plan generation for adaptive scaffolding in game-based learning environments. Int. J. Artif. Intell. Educ. (2024). https://doi.org/10.1007/s40593-024-00421-1
10. Karkalas, S., Chalvatza, F., Mavrikis, M., Nikolaou, M.S.: Using the repertory grid technique in a co-design process for learning analytics. In: 14th International Conference on Methodologies and Intelligent Systems for Technology Enhanced Learning, pp. 325–337. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-73538-7_29
11. Lallé, S., Yalcin, N., Conati, C.: Combining data-driven models and expert knowledge for personalized support to foster computational thinking skills. In: Proceedings of the 11th ACM International Conference on Learning Analytics & Knowledge, pp. 375–385. ACM, Irvine, CA, USA (2021)
12. Mavrikis, M., Karkalas, S., Cukurova, M., Papapesiou, E.: Participatory design to lower the threshold for intelligent support authoring. In: Artificial Intelligence in Education, pp. 185–189. LNCS, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23207-8_35
13. Saba, J., Kapur, M., Roll, I.: The development of multivariable causality strategy: instruction or simulation first? In: Artificial Intelligence in Education, pp. 41–53. Springer (2023). https://doi.org/10.1007/978-3-031-36272-9_4
14. Sabri, S., et al.: Integrating human-machine systems and digital twin technologies: navigating trust, interoperability, and ethical challenges. Interoperability Ethical Challenges (2025). https://doi.org/10.2139/ssrn.5189676
15. Saleh, A., Yuxin, C., Hmelo-Silver, C.E., Glazewski, K.D., Mott, B.W., Lester, J.C.: Coordinating scaffolds for collaborative inquiry in a game-based learning environment. J. Res. Sci. Teach. **57**(9), 1490–1518 (2020)
16. Shved, E., Bumbacher, E., Mejia-Domenzain, P., Kapur, M., Käser, T.: Teaching and measuring multidimensional inquiry skills using interactive simulations. In: Artificial Intelligence in Education, pp. 482–496. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64302-6_34
17. Sobocinski, M., Dever, D., Wiedbusch, M., Mubarak, F., Azevedo, R., Järvelä, S.: Capturing selfregulated learning processes in virtual reality: causal sequencing of multimodal data. Br. J. Edu. Technol. **55**(4), 1486–1506 (2024). https://doi.org/10.1111/bjet.13393
18. Swamy, V., Radmehr, B., Krco, N., Marras, M., Käser, T.: Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCs. arXiv:2207.00551 (2022)
19. Swamy, V., Romano, D., Desikan, B.S., Camburu, O.M., Käser, T.: illuminate: an LLM-XAI framework leveraging social science explanation theories towards actionable student performance feedback. arXiv:2409.08027 (2025)
20. Wang, K.D., Cock, J.M., Käser, T., Bumbacher, E.: A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. Br. J. Edu. Technol. **54**(1), 192–221 (2023). https://doi.org/10.1111/bjet.13289
21. Zhang, N., Biswas, G., Hutchins, N.: Measuring and analyzing students' strategic learning behaviors in open-ended learning environments. Int. J. Artif. Intell. Educ. 1–40 (2021). https://doi.org/10.1007/s40593-021-00275-x

Applications of Generative AI to Support Teaching and Learning in Higher Education: A Half-Day Workshop

René F. Kizilcec[1(✉)], Ryan S. Baker[2], Christopher Brooks[3], Jadon Geathers[1], Yann Hicke[1], Steven Moore[4], and Bo Wu[5]

[1] Cornell University, Ithaca, NY 14853, USA
{kizilcec,jag569,ylh8}@cornell.edu
[2] University of Pennsylvania, Philadelphia, PA 19104, USA
rybaker@upenn.edu
[3] University of Michigan, Ann Arbor, MI 48109, USA
brooksch@umich.edu
[4] Carnegie Mellon University, Pittsburgh, PA 15213, USA
stevenmo@andrew.cmu.edu
[5] Colorado School of Mines, Golden, CO 80401, USA
bwu@mines.edu

Abstract. Generative AI-based technologies offer new ways to engage students in personalized, inclusive, and interactive learning experiences. This includes applications such as course assistants, simulations, interactive activities, and tailored feedback on assignments. The goal of this half-day workshop is to share ongoing work in this rapidly evolving space and discuss what features of implementations contribute to successful adoption and use in university courses. Bringing together researchers, educators, and developers, we will showcase cutting-edge systems and empirical findings that examine the effectiveness of AI in higher education. Through presentations and group activities, participants will collaboratively identify emerging trends, challenges, and best practices. The workshop aims to synthesize insights into a set of recommendations to guide future research and implementation of AI-enhanced teaching and learning strategies.

Keywords: Higher education · Generative AI · Pedagogy · Teaching and Learning

1 Motivation, Content, and Themes

Over the past three years, interest in generative AI (GenAI) at the AIED conference has grown exponentially, reflecting the emergence of innovative approaches to enhance teaching and learning. LLMs in particular open new opportunities to address ongoing challenges in this space and advance higher education practices. While many of these interventions benefit education broadly, there remains a

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 267–274, 2025.
https://doi.org/10.1007/978-3-031-99267-4_34

pressing need to examine their practical implications specifically within universities. Factors such as learner maturity, class sizes, subject areas, and instructor experience all play pivotal roles in shaping how GenAI can be most effectively deployed in this setting.

This half-day workshop therefore centers on the application of GenAI for supporting teaching and student learning in higher education. We aim to showcase how to develop and use GenAI interventions that empower instructors and students alike. Participants will learn strategies to successfully implement GenAI in their courses, engage in hands-on activities with relevant tools, and explore the broader opportunities that ongoing advancements in GenAI afford. We invite researchers, educators, and practitioners eager to expand and refine teaching and learning practices at the university level to join us. It will focus on exploring the opportunities and challenges of integrating GenAI in supporting teaching and learning in higher education teaching. Key topics and themes include:

AI Course Assistants – The use of AI-powered assistants or chatbots to support instructors and students in courses [3,5,8]. For example, AI teaching assistants can answer routine questions on discussion forums, as demonstrated by Georgia Tech's Jill Watson (an IBM Watson-based TA that handled forum Q&A in an online class) [2]. Such AI assistants can scale support for large classes, freeing human instructors for complex tasks. Notably, early trials showed students often didn't realize a responder was an AI, indicating these systems can effectively mimic helpful human support.

Feedback Generation and Automated Assessment – Techniques for leveraging GenAI to provide personalized feedback on student work. Large language models (LLMs) like Anthropic's Claude Sonnet, Google's Gemini 2.0, and OpenAI's GPT-4o can be used to generate formative feedback on essays, projects, or coding assignments. Recent research has explored using LLMs to write human-like feedback that is informative and encouraging for students [6,9]. This theme examines how AI-generated feedback might improve learning by giving students timely, tailored guidance, as well as how instructors can curate or edit AI feedback to ensure quality. We will discuss scenarios such as AI tools that summarize a student's work and offer suggestions, helping instructors handle large classes or provide 24/7 support.

AI-Integrated Learning Activities – Design of learning activities and assignments that meaningfully incorporate AI tools. Rather than banning AI, some educators now require students to use AI for certain tasks to enhance learning. We will share use-cases like AI-assisted writing exercises, using AI to generate quiz questions or study guides, and creative projects where students collaborate with AI (e.g. using an AI as a brainstorming partner or a coding helper). This theme also covers how AI can personalize learning experiences by adapting content or practice questions to individual student needs, aligning with the goal of more inclusive and personalized education [11].

Challenges in Adoption – A critical look at the hurdles institutions and instructors face when incorporating GenAI into teaching. A primary concern is the risk of students using AI to circumvent learning (e.g. getting answers or

writing from AI without understanding) [7]. We will discuss academic integrity issues and strategies to prevent misuse, such as designing "AI-proof" assessments or teaching AI literacy so students use GenAI responsibly. Another challenge is faculty preparedness and training not all educators are comfortable with AI, and many need guidance on effective integration. Technical limitations of current GenAI will be addressed as well, including the well-known tendency of LLMs to "hallucinate" (make up plausible-sounding but incorrect information), and how this impacts trust and accuracy in educational use. Participants will collaboratively consider solutions to these challenges.

Ethical and Inclusiveness Considerations – Throughout all topics, we will emphasize the ethical implications and the goal of equitable, inclusive education. This includes ensuring AI tools are fair and unbiased, respect student privacy, and are accessible to diverse student populations, which is a known challenge in this space [1,4]. The workshop will explore questions like: How do we mitigate AI biases that could disadvantage or offend certain groups? How can GenAI be used to reduce barriers for marginalized students (for instance, providing extra tutoring support at no cost, or offering translations and adaptive content)? We will invoke guidelines on responsible AI in education, aligning with AIED 2025s theme of using AI to empower teachers and students for an equitable future. Case studies of ethical AI usage and policies (such as disclosure of AI use, data transparency, and obtaining consent for AI analysis of student data) will be discussed. Ultimately, this theme is about ensuring that GenAI is a positive catalyst for inclusion and ethics in the classroom rather than a source of new inequalities [10].

1.1 Significance and Relevance to the AIED Community

As GenAI continues to reshape education, university instructors and researchers are grappling with its potential to enhance learning while ensuring ethical and equitable integration. This workshop aligns with the AIED 2025 theme, *"AI as a Catalyst for Inclusive, Personalised, and Ethical Education,"* by examining how GenAI can be leveraged to empower both teachers and students in higher education.

In this workshop, we will discuss how AI can personalize learning, address diverse student needs, and reduce educational disparities. A key focus will be on ensuring ethical implementation, emphasizing fairness, transparency, and privacy in AI-based tools. By bringing together educators, researchers, and developers, this workshop will foster collaborative problem-solving on the challenges and opportunities of AI adoption in higher education, ensuring that these technologies serve as enablers of inclusive and equitable education rather than sources of bias or exclusion.

This workshop will contribute to the AIED community by synthesizing best practices, empirical insights, and future research directions for AI-driven teaching and learning. It will provide a critical forum for discussing how AI can empower instructors, offer personalized support to students, and shape the next generation of ethical AI-enhanced education.

2 Proposed Timeline and Process

The call for papers will open on March 27, 2025, and the submission deadline will be May 29, 2025. The period from April through early May 2025 will be dedicated to workshop publicity and submission collection. Interested participants can submit **short papers** (4–6 pages) or **extended abstracts** (1–2 pages) describing their research, experience reports, demonstrations, or position statements related to GenAI in higher education teaching. The submission window of 6–8 weeks should provide authors with adequate preparation time. Submissions will be managed through EasyChair (or another system provided by AIED 2025). Our workshop program committee will conduct a thorough but prompt review process (approximately 2 weeks) following the submission deadline. Three members of the program committee will review each submitted paper. Authors will be notified of their paper acceptance on June 12, 2025. Accepted contributions may include papers for oral presentation, poster/demo descriptions, and lightning-talk proposals. Camera-ready versions will be requested by early July for any planned proceedings or online materials. In the weeks preceding the conference, specifically early July 2025, we will finalize the workshop schedule, organize accepted talks into thematic sessions, and confirm any invited speakers or panelists. The complete program will be published on the workshop website by early July and circulated to all registered participants along with any pre-reading materials.

We will actively **advertise the workshop** while accepting contributions. We will promote it through targeted email campaigns to the AIED community mailing list and related interest groups. Our team members will share the call for participation within their respective institutional networks and with colleagues at universities with strong AI in education research programs. All promotional materials will include the workshop's key themes, submission deadlines, and a link to the detailed workshop website.

3 Format and Activities

The workshop will follow an interactive and structured format designed to foster engagement and knowledge sharing among participants. As outlined in Table 1, the half-day event will include a combination of presentations, discussions, and hands-on group activities. This will be followed by a panel discussion on challenges, ethics, and implementation, addressing critical issues such as responsible AI use and academic integrity. The final session will involve an interactive group activity, where participants collaboratively design AI-powered interventions for education. The workshop will conclude with closing remarks and a discussion on future collaborations, ensuring that participants leave with actionable insights and connections for continued engagement.

Table 1. Half-day Workshop Schedule

Time	Session
09:00 - 09:20	**Introduction and Welcome** Overview of workshop objectives and introduction of organizers and attendees.
09:20 - 10:20	**Session 1: Presentations on Cutting-Edge Research** Selected short paper presentations highlighting recent research on AI course assistants, AI-generated feedback, and AI-integrated learning activities. Each talk will focus on the *design, implementation, and empirical evaluation* of AI-based educational tools. **Q&A and discussion after each presentation.**
10:20 - 10:35	**Break**
10:35 - 11:20	**Session 2: Challenges, Ethics, and Practical Implementation** Lightning talks and panel discussion featuring experts on: – Ensuring responsible and ethical AI usage in classrooms. – Institutional challenges in adopting AI tools for education. – Addressing student misuse, academic integrity, and faculty training. **Panel Q&A with audience participation.**
11:20 - 12:05	**Session 3: Interactive Group Activity Designing AI-Integrated Learning Strategies** Participants will be divided into small groups to discuss and prototype AI-based interventions for their courses or research projects. Groups will explore: – How to design *effective AI-driven learning activities*. – Best practices for *AI-generated feedback* and student engagement. – Identifying potential risks and mitigation strategies. Each group will present their ideas in a **brief share-out session**.
12:05 - 12:20	**Closing Remarks and Next Steps** Summary of key takeaways, opportunities for future collaborations, and call for continued engagement.

4 People

4.1 Program Committee

The following researchers (in alphabetical order) will serve as the Workshop Program Committee to help shape the technical program:

– **Ryan Baker** is a Professor at the University of Pennsylvania and the Director of the Penn Center for Learning Analytics, specializing in Educational Data Mining and Learning Analytics. His research focuses on AI-driven models to analyze student learning behaviors and improve adaptive learning systems in digital education.

- **Christopher Brooks** is an Associate Professor at the University of Michigan and is an applied Computer Scientist who builds and studies the effects of educational technologies in higher education and informal learning environments.
- **Jadon Geathers** is a PhD student in Information Science at Cornell University, advised by Rene Kizilcec. His research examines how conversational AI can transform educational technologies through scalable practice opportunities, interactive dialogue, and automated feedback systems.
- **Yann Hicke** is a PhD student in Computer Science at Cornell University, advised by Claire Cardie and Rene Kizilcec. His research focuses on AI-driven conversational simulations, multimodal dialogue analysis, and adaptive feedback generation for training and assessment.
- **Rene Kizilcec** is an Associate Professor of Information Science at Cornell University, where he directs the Future of Learning Lab. He studies the design and deployment of AI in higher education, including AI-powered course assistants and simulation-based learning platforms.
- **Steven Moore** is a postdoctoral researcher at the Human-Computer Interaction Institute in the School of Computer Science at Carnegie Mellon University, advised by Dr. Ken Koedinger. His research is on learnersourcing, crowdsourcing, and human-AI partnerships.
- **Jiliang Tang** is a University Foundation Professor in the Computer Science and Engineering Department at Michigan State University. His research interests include graph machine learning, trustworthy AI, and their applications in biology and education.
- **Bo Wu** is an Associate Professor in the Department of Computer Science at Colorado School of Mines. He works on building efficient systems for machine learning and graph processing applications. He is the CEO and co-founder of HITA AI, an education AI platform for educators and students.

4.2 Potential Authors

We anticipate strong interest from researchers, educators, and developers working at the intersection of generative AI and higher education. Potential authors include those conducting empirical research on AI-powered teaching and learning, designing and evaluating AI course assistants, developing AI-driven feedback systems, and implementing AI-integrated learning activities. Thus, contributions may come from:

- **Researchers** in Artificial Intelligence in Education (AIED), Learning Analytics (LAK), Learning at Scale (L@S), Educational Data Mining (EDM), Human-Computer Interaction (HCI), and Computer-Supported Collaborative Learning (CSCL), including researchers who study the pedagogical impact, ethical considerations, and effectiveness of AI-based tools in university settings.

- **Developers** and **Industry Researchers** who design and build AI-driven educational tools and who have insights into technical design, usability, and implementation challenges.
- **Policy and Ethics Experts** working on the responsible use of AI in education, ensuring that AI-based tools are fair, inclusive, transparent, and aligned with academic integrity standards.
- **Instructors** experimenting with AI in the classroom, who can share case studies on successful (and unsuccessful) AI integration into courses.

4.3 Workshop Audience

Beyond the program committee and potential authors, we anticipate interest in the workshop from members of the AIED, EDM, and L@S communities. The anticipated maximum number of participants is 40 people.

References

1. Dotan, R., Parker, L.S., Radzilowicz, J.: Responsible adoption of generative AI in higher education: Developing a "points to consider" approach based on faculty perspectives. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 2033–2046 (2024)
2. Goel, A.K., Polepeddi, L.: Jill Watson: a virtual teaching assistant for online education. In: Learning Engineering for Online Education, pp. 120–143. Routledge (2018)
3. Hicke, Y., Agarwal, A., Ma, Q., Denny, P.: AI-TA: towards an intelligent question-answer teaching assistant using open-source LLMs. arXiv preprint arXiv:2311.02775 (2023)
4. Lee, J., Hicke, Y., Yu, R., Brooks, C., Kizilcec, R.F.: The life cycle of large language models in education: a framework for understanding sources of bias. Br. J. Edu. Technol. **55**(5), 1982–2002 (2024)
5. Liu, C., Hoang, L., Stolman, A., Wu, B.: HiTA: A RAG-based educational platform that centers educators in the instructional loop. In: International Conference on Artificial Intelligence in Education, pp. 405–412. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64299-9_37
6. 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)
7. Pardos, Z.A., Bhandari, S.: ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. PLoS ONE **19**(5), e0304013 (2024)
8. Shah, M., Pankiewicz, M., Baker, R.S., Chi, J., Xin, Y., Shah, H., Fonseca, D.: Students' use of an LLM-powered virtual teaching assistant for recommending educational applications of games. In: Joint International Conference on Serious Games, pp. 19–24. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-74138-8_2
9. Stamper, J., Xiao, R., Hou, X.: Enhancing LLM-based feedback: Insights from intelligent tutoring systems and the learning sciences. In: International Conference on Artificial Intelligence in Education, pp. 32–43. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64315-6_3

10. Viberg, O., Kizilcec, R.F., Wise, A.F., Jivet, I., Nixon, N.: Advancing equity and inclusion in educational practices with AI-powered educational decision support systems (AI-EDSS) (2024)
11. Yan, L., Greiff, S., Teuber, Z., Gašević, D.: Promises and challenges of generative artificial intelligence for human learning. Nat. Hum. Behav. **8**(10), 1839–1850 (2024)

Intelligent Textbooks: The Sixth International Workshop

Sergey Sosnovsky[1], Peter Brusilovsky[2(✉)], Andrew Lan[3], and Isaac Alpizar-Chacon[1]

[1] Utrecht University, Princetonplein 5, 3584 CC Utrecht, Netherlands
{s.a.sosnovsky,i.alpizarchacon}@uu.nl
[2] University of Pittsburgh, 135 North Bellefield ave., Pittsburgh, PA 15260, USA
peterb@pitt.edu
[3] University of Massachusetts Amherst, Amherst, MA 01003, USA
andrewlan@cs.umass.edu

Abstract. From the early days of digital textbooks to the rapidly progressing age of Large Language Models, researchers from different areas explored multiple research directors on the crossroads of Textbooks, a traditional learning medium and Artificial Intelligence, a technology that could empower it. These research directions formed a new research area, often referred to as *Intelligent Textbooks*. The International Workshop on Intelligent Textbooks at AIED 2025, the sixth workshop in the series, aims to bring together researchers working on different aspects of Intelligent Textbooks to exchange complementary insights, review new results, and discuss emerging ideas.

Keywords: Online Textbooks · Adaptive Textbooks · Intelligent Textbooks

1 Introduction and Background

Textbooks have evolved over the last several decades in many aspects. Most textbooks can be accessed online, many of them freely. They often come with libraries of supplementary educational resources or online educational services built on top of them. As a result of these enrichments, new research challenges and opportunities emerge that call for the application of AIED methods to improve digital textbooks, their development, and the interaction of learners with them. This workshop seeks research contributions addressing these and other research questions related to the idea of intelligent textbooks. The workshop aims at bringing together researchers working on different aspects of learning technologies and AI to advance intelligent textbooks as a new interdisciplinary research field. The workshop builds on the success of the five previous events [2–6] and the recent special issue on Intelligent Textbooks published by the International Journal of AI in Education [7].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 275–277, 2025.
https://doi.org/10.1007/978-3-031-99267-4_35

Several exciting new developments have been reported since the last workshop, which was organized in 2023. In particular, we expect that the iTextbooks 2025 workshop will reflect on the potential of generative AI techniques in intelligent textbooks. Many recent works on intelligent textbooks incorporate these technologies, especially large language models (LLMs). Digital textbooks have contributed greatly to the pre-training of LLMs, providing them with information on the knowledge and structure of textbook content. This opens many pathways to explore the synergy between LLMs and textbooks to enhance the functionality of textbooks with intelligent and adaptive services and enrich interaction between textbooks and their readers.

2 Topics of Interest

A review of 30 years of research on intelligent textbooks [1,7] revealed several broad research areas that address various aspects of making textbooks more intelligent and using artificial intelligence to develop a new generation of textbooks. This year, the workshop accepted contributions from all these areas as well as on a set of recently emerged research topics. More specifically, we focus on the following groups of topics:

- Modeling and representation of textbooks: examining the prerequisite and the semantic structure of textbooks to enhance their readability;
- Analysis and mining of textbook usage logs: analyzing the patterns of learners' use of textbooks to obtain insights on learning and the pedagogical value of textbook content;
- Generation, manipulation, and presentation: exploring and testing different formats and forms of textbook content to find the most effective means of presenting different knowledge;
- Assessment and personalization: developing methods that can generate assessments and enhance textbooks with adaptive support to meet the needs of every learner using the textbook;
- Knowledge visualization: augmenting textbooks with concept maps, open learner models and other knowledge-rich extensions;
- Collaborative technologies: building and deploying social components of digital textbooks that enable learners to interact with not only content but other learners;
- Smart interactive content: extending online textbooks with various kinds of smart interactive content to improve learning, engagement, learner modeling, and personalization;
- Intelligent information retrieval and question-answering for digital textbooks;
- Content curation and enrichment: sorting through external resources on the web and finding the relevant resources to augment the textbook and provide additional information for learners.

Acknowledgments. The workshop is partially supported by the National Science Foundation through the SPLICE Infrastructure project.

References

1. Brusilovsky, P., Sosnovsky, S., Thaker, K.: The return of intelligent textbooks. AI Mag. **43**(3), 337–340 (2022)
2. Sosnovsky, S., Brusilovsky, P., Baraniuk, R., Agrawal, R., Lan, A. (eds.): Proceedings of the First International Workshop on Intelligent Textbooks. CEUR-WS (2019). https://ceur-ws.org/Vol-2384/
3. Sosnovsky, S., Brusilovsky, P., Baraniuk, R., Lan, A. (eds.): Proceedings of the Second International Workshop on Intelligent Textbooks. CEUR-WS (2020). https://ceur-ws.org/Vol-2674/
4. Sosnovsky, S., Brusilovsky, P., Baraniuk, R., Lan, A. (eds.): Proceedings of the Third International Workshop on Intelligent Textbooks. CEUR-WS (2021). https://ceur-ws.org/Vol-2895/
5. Sosnovsky, S., Brusilovsky, P., Lan, A. (eds.): Proceedings of the Fourth International Workshop on Intelligent Textbooks. CEUR-WS (2022). https://ceur-ws.org/Vol-3192/
6. Sosnovsky, S., Brusilovsky, P., Lan, A. (eds.): Proceedings of the Fifth International Workshop on Intelligent Textbooks. CEUR-WS (2023). https://ceur-ws.org/Vol-3444/
7. Sosnovsky, S., Brusilovsky, P., Lan, A.S.: Intelligent textbooks. Int. J. Artif. Intell. Educ. (2025). https://doi.org/10.1007/s40593-024-00451-9

Full-Day Tutorial: Conducting in-the-Moment Data-Driven Classroom Interviews (DDCI) During Digital Learning Using Quick Red Fox (QRF)

Jaclyn Ocumpaugh[1], Luc Paquette[2(✉)], Amanda Barany[1], and Ryan Baker[1]

[1] University of Pennsylvania, Philadelphia, PA, USA
[2] University of Illinois Urbana-Champaign, Champaign, IL, USA
lpaq@illinois.edu

Abstract. This tutorial will introduce participants to Detector-Driven Classroom Interview (DDCI) research. It will demonstrate how to conduct data-driven interviews using the Quick Red Fox (QRF) application through a series of hands-on activities. Participants will be able to attend the qualitative research track, where they will develop and practice interview techniques to be used during classroom interviewing, or to attend the computational track, where they will practice developing simple Python scripts that can be used to trigger an interview request on the QRF application.

Keywords: Data-driven interviews · digital learning environment · mixed-method research

1 Organizers

Ocumpaugh is the Associate Director of the Penn Center for Learning Analytics at the University of Pennsylvania. Before working with Paquette, Baker and colleagues to invent the methodology for Data Driven Classroom Interviews (DDCI), she was one of the co-creators of the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) for studying student engagement. In addition to her university teaching experience, she has extensive experience running field-training in observation methods and has already trained several interviewers in the DDCI methodology.

Paquette is an Associate Professor in the department of Curriculum & Instruction at the University of Illinois Urbana-Champaign. Paquette co-developed the QRF technology and the DDCI methodology with Ocumpaugh, Baker and colleagues. His research focuses on the development and application of computational methods to study log trace data produced by students' interactions with digital learning environments. He has prior experience with organizing workshops at international research conferences (Learning Analytics and Knowledge, and The International Educational Data Mining Conference).

Barany is a Post Doc at the Penn Center for Learning Analytics with substantial experience researching digital learning environments, including conducting DDCIs. She

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 278–283, 2025.
https://doi.org/10.1007/978-3-031-99267-4_36

brings a substantial amount of teaching and leadership experience to this project, as well as research experience in coding and analyzing DDCI transcripts.

Baker is a Professor in the Graduate School of Education at the University of Pennsylvania and director of the Penn Center for Learning Analytics. Baker helped to co-develop the QRF technology and associated DDCI methodology and has extensive experience in both education research and field training methods, as well as a long history of organizing research meetings and events.

2 Overview

The goal of this tutorial is to provide an introduction to Data Driven Classroom Interviews (DDCIs), a new methodology that has been developed by our research team. DDCIs can be conducted using an Android application called Quick Red Fox (QRF), which pings a researcher in the classroom when an automated model detects a construct, behavior, or event of interest. The researcher can then conduct an immediate interview, while the student is experiencing that event, leveraging the benefits of a talk-aloud interview while providing additional metacognitive support and research direction that allows us to get beyond the ideas or feelings that are immediately salient to the student. This is a considerable advance over talk-aloud, as these data-driven classroom interviews (DDCIs) are less disruptive to a normal classroom (allowing greater ecological validity than a laboratory study). They also allow greater ecological validity than retrospective focus groups, where students who have since resolved any struggle may report the experience differently than students who have not.

To date, there are two primary methods for obtaining real-time data about students' learning. One is to run various kinds of classroom observations, often with limited interactions between the researcher and the students. The other is the Think Aloud (TA) method, a technique grounded in information processing theory, where students are asked to verbalize their cognitive processes [1]. The latter technique facilitates significantly more access to students' internal thoughts, but it has been criticized for increasing the cognitive load of the activity and for being difficult to implement in User Experience (UX) Design situations which cause frustration [2], which may be the exact time that researchers studying digital learning would like to hear the most from a given student. Moreover, it can be difficult to conduct in a regular classroom. Still, TA was created over concerns about alternative methods, including the ways in which memory–even when it is functioning ultimately–consolidates and clouds people's interpretation of events [1]. These concerns mean that TA can be preferable to alternatives like retrospective interviews.

DDCIs are meant to address some of the methodological concerns related to the TA approach while facilitating real-time access to students cognitive, affective, and metacognitive processes. By having a researcher direct questions towards students at optimal times, we offer several advantages over TA, including (a) we are less likely to either interfere with the totality of the learning process and/or to risk fatigue because we are only asking students to talk to us at specific times, (b) we are not asking them to read our minds (i.e. to figure out what information about their experience is relevant to our research questions), which subsequently (c) reduces the cognitive burden of the research

process. That is, if a student's verbalizations are incomplete, the researcher is there to ask follow-up questions that help us to distinguish between the student not having any information to share about their experience and the student not remembering to explain those details while also trying to learn.

Moreover, it increases the flexibility of the research process, as research subjects often require significant training to be able to understand what a researcher needs them to do during a TA event [3]. In contrast, DDCIs allow the researcher to modify the questions that students are trying to answer during the research process. This flexibility has allowed us to increase the range of questions that we asked about students' self-regulated learning (as in [4]), but it has also facilitated rapid design changes to a pedagogical agent that substantially improved students' interactions with that system [5].

That said, the process of connecting with students across repeated interviews requires careful planning and practice. Researchers must identify appropriate times to conduct interviews (designing triggers for the QRF app), must design these triggers in a way that does not interrupt a given student too frequently, and then must execute these interviews in a way that builds trust with students while also asking interview questions that reflect their own research goals. In this tutorial, we will cover strategies required for all stages of a DDCI project, providing training to participants who are interested in conducting this–or similar research–on their own.

3 Outline of Planned Activities

The workshop will be broken down into several smaller, interactive sessions. Morning activities will introduce the DDCI approach, provide a live demonstration of its usage and concrete examples of research projects where the method was applied. In addition, participants will be engaged in discussions about how they could apply the DDCI method in the context of their own research.

Afternoon activities will engage participants in hands-on training in using the DDCI method. This will include looking at, discussing and analyzing sample interview data collected using the QRF application. Participants will separate into two tracks. Participants in the Qualitative Track will further develop ideas about how they would analyze interview data collected using the QRF application. Participants in the Computational Track will learn technical information about the QRF infrastructure. This will include information about how the QRF application and server communicate with each other. Participants will also learn how to write simple Python scripts that can trigger an interview request on the QRF application (Table 1).

4 Intended Audience

This tutorial is targeted towards participants who are working with digital learning environments that are capable of processing students' data in real time. However, we are very receptive to inviting participants who are still in the very early planning stages of doing so. Recruitment will include advertisements to several listservs that target researchers in the learning sciences, learning analytics, AIED, and educational data mining communities. Ocumpaugh, Paquette, Barany and Baker have extensive research

histories in these communities which should help to increase the reach of these efforts. We will have designed this tutorial to accommodate 8-20 participants, though this could be expanded relatively easily if there is increased interest.

Table 1. Proposed timeline (flexible to accommodate the needs of break/meals scheduling).

Time	Activity	Description
9:00	Introductions	Who are you and what are your research questions? What do you need to know about your students' interactions with an digital learning system
9:30	Intro to QRF	Interactive Presentation: High-level overview of how the tool works. Workshop users will be able to download the app and/or borrow a phone to see how it works.
10:00	DDCI Overview	High-level overview of strategies, research, and how DDCIs compare to talk-aloud or retrospective interview methods. Overview of different kinds of interview strategies and building connections with students.
10:30	Break	
10:45	Refining your research questions	Paired activity: What research questions does your team have that could be helped by contextualizing your student's process log data? What are students able to articulate about their experience that might inform your analyses?
11:30	Lunch	
12:00	Intro to Analyzing Interview Data	Interactive: What kinds of analyses do you want to run on your data? This will help to guide how frequently you need to sample each student and the kinds of questions you'll need to ask.
12:30	Drafting your interview questions	Paired activity: What kinds of interview questions should you use as you are getting to know the students? What kinds of interactions build trust? What kinds of interview questions will help you to answer your research questions?

(*continued*)

Table 1. (*continued*)

Time	Activity	Description
1:00	Workshopping interview questions	Group activity: What common techniques/questions can we see across groups? How can we make sure that we are identifying students' strengths (and not just their points of struggle). This will also allow participants to volunteer to role-play, with some acting as students (playing an educational game) and others acting as interviewers.
1:30	Break	
1:45	When should we interview?	Paired activity: How do we decide when to conduct interviews. For some constructs, it is probably relatively easy, but others may be more difficult. Do we need control groups? What themes emerged around trigger design? How do these themes relate to the role-playing exercises from earlier?
2:05	Break	
Break-out Session A: Qualitative Track		
2:30	Analysis Design Overview	How do you identify themes in your data? When do you need to achieve inter-rater reliability? We will provide sample transcripts from student interviews and talk about the qualitative coding processes as well as potential subsequent analyses.
3:00	Analysis Design Activity	Paired activity: Use additional transcripts (we will provide) to identify themes related to your own research questions
Breal-out Session B: Computational Track		
2:30	Overview of writing trigger code	Technical overview of how the QRF app and server work together. Overview of develop scripts that process your data and trigger interview requests (for those who already have python experience)
3:00	Trigger code writing activity	Python activity where tutorial participants practice writing code that would facilitate a QRF/interview trigger in an educational game.

(*continued*)

Table 1. (*continued*)

Time	Activity	Description
4:00	Break	
4:15	Wrap-up and evaluation	What additional questions/concerns do you have? What feedback do you have for us on this tutorial?

5 Previous Editions

While this tutorial has never been conducted yet, the same tutorial proposal was submitted and accepted for the Annual Meeting of the International Society of the Learning Sciences (ISLS), which will take place from June 10-13.

Acknowledgements. This work was supported by the National Science Foundation (NSF; ECR-2301172 and ECR-2301173). Any conclusions expressed in this material do not necessarily reflect the views of the NSF.

References

1. Ericsson, K.A., Simon, H.A.: Verbal reports as data. Psychol. Rev. **87**(3), 215 (1980)
2. Boren, T., Ramey, J.: Thinking aloud: reconciling theory and practice. IEEE Trans. Prof. Commun. **43**(3), 261–278 (2000)
3. Whitehead, A.E., et al.: Changes in cognition over a 16.1 km cycling time trial using think aloud protocol: preliminary evidence. Int. J. Sport Exerc. Psychol. **17**(3), 266-274. (2019)
4. Andres, J.M.A.L., Hutt, S., Ocumpaugh, J., Baker, R.S., Nasiar, N., Porter, C.: How anxiety affects affect: a quantitative ethnographic investigation using affect detectors and data-targeted interviews. In: Proceedings of the International Conference on Quantitative Ethnography (2021).
5. Ocumpaugh, J., et al.: Using qualitative data from targeted interviews to inform rapid AIED development. In: Proceedings of the 29th International Conference on Computers in Education, pp. 69–74 (2021)

Ethical Artificial Intelligence and Education: The Need for International Regulation to Foster Human Rights, Democracy and Equity

Christian M. Stracke[1,5(✉)], Beth Havinga[2,5], Wayne Holmes[3,5], and Ron Salaj[4,5]

[1] University of Bonn, Bonn, Germany
stracke@uni-bonn.de
[2] European EdTech Alliance, Marienfeld, Germany
[3] University College London, London, UK
[4] University of Torino, Turin, Italy
[5] Council of Europe AI&ED Expert Group, Strasbourg, France

Abstract. The growing implementation of Artificial Intelligence in education (AIED) raises pressing ethical, legal, and political questions. Unlike other sectors, education affects minors and youth, shapes societal values and civic participation, and plays a foundational role in upholding democratic principles. Education cannot, therefore, be treated as a standard application domain for AI technologies. The introduction of AI systems in educational environments risks reinforcing inequality, diminishing democratic oversight, and displacing core human functions in pedagogy. This paper introduces the concept of the workshop held at the international 26th AIED Conference 2025. It responds to these concerns by focusing on the development of binding international regulations and organisational safeguards to govern the complex relationship between AI and education (AI&ED). Grounded in the work of the Council of Europe's AI&ED Expert Group, it aims to identify specific legal requirements and structural measures necessary to protect human rights, democracy, and the rule of the law in educational contexts. Through interactive formats, the workshop will critically examine existing policy proposals, identify omissions, and propose principles for effective regulation that serve the common good. The results can feed into the Council of Europe's development of two international instruments: a binding convention on AI in education and a recommendation on AI literacy.

Keywords: Artificial Intelligence and Education (AI&ED) · regulation · human rights · democracy · Council of Europe · digital equity

1 Introduction

The adoption of Artificial Intelligence (AI) technologies in education is proceeding rapidly, often with minimal public debate or sufficient regulatory scrutiny [5, 6, 20, 21]. This expansion brings with it not only technical challenges, but also ethical and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 284–291, 2025.
https://doi.org/10.1007/978-3-031-99267-4_37

societal concerns [16, 19]. The domain of education is distinct from other sectors in that it deals with users of systems who have limited agency over the tools that they use and exist within an environment that is predominantly mandatory [15, 18]. Education environments play a vital role in the safeguarding and enactment of societal values and development of democratic awareness and principles [13, 15].

Unregulated AI use in education risks exacerbating social inequalities, displacing critical pedagogical roles, undermining essential principles of democracy and further limiting the agency of learners and educators with regard to safe data practices [4, 14]. Standard regulatory approaches that treat AI as a neutral tool are insufficient to address the power asymmetries and institutional vulnerabilities specific to education [15].

This context demands a targeted, rights-based regulatory framework that prioritises transparency, public accountability, and democratic control.

1.1 Background

In 2021, the Council of Europe (COE) established an expert group exploring the application of AI in education (AIED), the teaching of AI in education (AI Literacy) and the regulatory needs associated with these areas [7]. The goal of the group is to develop actionable insights and recommendation for CoE Member States, which cover 46 countries within wider Europe and are responsible for a population of approximately 675 million people. These recommendations are intended to guide the application of AI in education for the common good and aligned with the essential focus areas of the CoE: Human rights, democracy and the rule of the law [8].

In 2022, the CoE expert group on education prepared the report, *Artificial Intelligence and Education: A critical view through the lens of human rights, democracy and the rule of law* [11]. This report highlighted the need for education-specific safeguarding, AI governance structures, and was designed to further critical discourse. The report served as an important step towards developing two policy instruments as requested by the Member States: a binding legal instrument regulating the use of AI in education, and a recommendation covering AI literacy across educational systems specifically focusing on human-centred approaches [7]. After receiving a mandate in September 2023 from all Member States, work began on drafting these international instruments [8].

This workshop provides an opportunity for critical scrutiny and stakeholder engagement to inform these developments and ensure their democratic legitimacy.

2 Workshop Objectives and Importance

2.1 Workshop Objectives

This workshop at the AIED 2025 in Palermo, Italy [3] is the continuation of the very successful discussions held at the AIED 2023 in Tokyo, Japan [1, 12] and AIED 2024 in Recife, Brazil [2, 17]. The aim of the workshop is to continue important global discourse on regulatory practices regarding AI and education whilst at the same time providing important and ongoing updates and raising awareness for the important topic of ethical AI&ED. Additionally, the workshop discussions are intended to contribute to the CoE's current developments of an international AI regulation in education as binding law in Europe and design of comprehensive AI literacy recommendations.

The workshop is designed to help participants:

- Think through what kinds of ethical and legal guidelines are necessary for the responsible use of AI in education and critically examine how these could be applied in real contexts.
- Discuss how AI might influence the future of education, including questions of trust, fairness, and responsibility, and bring together different perspectives to better understand the consequences.
- Work collaboratively to explore how rules and policies for ethical AI use can be created, focusing on what legal and institutional structures are needed and how they might be developed.
- Consider how to keep up with the rapid pace of AI development while still ensuring that regulation remains effective and meaningful over time.

These goals are tied to a set of guiding concerns:

- What steps are needed to put in place legal and organisational systems that support ethical use of AI in educational settings?
- In what ways can AI be introduced without weakening, but instead, and ideally, strengthening human rights, democracy, equal access to education, and the rule of law?
- And how can we make sure that stakeholders at all levels of the education system, from individual learners to national policymakers, are involved in shaping these decisions and ensuring AI is used in fair and transparent ways?

The main focus of the workshop will be on the legal and organizational requirements to achieve a regulation in the fields of AI and Education (AI&ED). The discussions will be based on the work and activities of the CoE and its appointed AI&ED expert group. In particular, the most recent work of this group regarding the preparation of the legal instrument and recommendation for AI literacy will form the basis of the workshop. In addition to a fundamental exploration of these themes, open and urgent issues will be addressed, and next steps in relation to these requirements and goals will be defined.

2.2 Workshop Relevance and Importance to the AIED Community

In preparation of the CoE Council of Ministers meeting in 2025 [7, 8], during which the proposed texts for the legal instruments on AI in education will be discussed and the final report from the expert group will be presented, the CoE expert group is currently collecting feedback. It is essential that this feedback reflects different stakeholder perspectives. These include those of students, teachers, industry representatives, NGOs, and academia.

The AIED research community plays an important role within this process, bringing valuable knowledge and experience to the discussions and ensuring that future legal instruments and guidelines reflect the realities of education. It is imperative that these perspectives are considered in the work of the CoE.

This workshop also gives members of the AIED community a chance to get directly involved and is closely aligned with the AIED 2025 conference theme, “AI as a Catalyst for Inclusive, Personalised, and Ethical Education”. During the workshop, participants

will learn about the work of the CoE that has already been conducted. This includes the results of the CoE AI&ED Expert conferences "Artificial Intelligence and Education" held on 21st / 22nd of November 2023 [4] and on 24th / 25th of October 2024 in Strasbourg [5]. Furthermore, the workshop will draw on the group's first report [11] in viewing AI in education from the perspective of human rights, democracy, and the rule of law and on the results from the two former AIED workshops [12, 17].

Through an active discussion, the workshop will also facilitate the formation of a wider community of practice within the AIED community with the goal of considering the wider societal implication of AIED.

3 Workshop Format and Methodology

This half-day workshop will be run in an interactive, hybrid format, encouraging active participation throughout. It is divided into two parts, each building on the other. The first part focuses on small group work, where participants will collaborate on four key questions.

- What assumptions or misrepresentations exist in the current CoE proposals?
- What issues have been neglected or insufficiently addressed?
- What types of legal and institutional mechanisms are needed to uphold rights?
- Which real-world examples best illustrate the risks and regulatory needs?

In the second part, the workshop will use the World Café approach, where participants can choose from up to ten prepared questions and explore them in rotating group discussions. The discussions will be documented, synthesised, and structured into categories that reflect actionable regulatory needs.

The workshop will take place with both in-person and online groups working in parallel. After each round of discussion, participants will come together in plenary sessions to share insights and reflect on the results. This format ensures that ideas and perspectives from all groups are exchanged and taken forward collectively.

3.1 Workshop Programme

In the first section, the interactive Workshop will begin with a brief introduction to the workshop and a keynote by the Council of Europe. It is followed by a summary of the key messages from the initial report on AI&ED and the current work by the CoE Expert Group (15 min).

The core part of the first section will centre on four key questions (the length will be dependent on the total time available). Each question will be introduced by a 5-min presentation. Participants will then work in small groups, to respond to the issues introduced, before reporting back their views and ideas to the whole workshop. The first section will conclude with a discussant summarising the workshop participants' input (15 min) followed by a wrap up (5 min).

In the second section, the interactive Workshop follows the World Café method and consists of four parts:

1. In the first short part (5 min), we will briefly present the World Café method and the topic "Requirements for international conventions to ensure ethical AI introduction and usage in education strengthening democracy building".
2. In the second core part (the length will be dependent on the total time available), we will discuss different and pre-prepared themes and aspects in groups of ideally 4 to 10 participants. We will prepare 10 themes before the conference and will decide depending on the number of participants how many group tables we will offer during the workshop. The procedure will be identical on all group tables:
   - The theme will be read that always consists of an open question and clarified.
   - Afterwards, the contributions (= answers on the open question of the theme) will be collected as post-it at a flipchart next to the group table.
   - Finally, all collected contributions will be sorted in open discussion and be clustered according defined categories or criteria so that they can be presented. That can be done independently or using the categories or criteria from the other rounds before.
3. In the third part (30 min), the results from the group tables will be presented in the plenary and critically discussed and reflected. These can then be provided as summaries for publication after the event.
4. In the fourth part, we will collect all issues that the workshop participants wish to discuss and bring to the attention of the CoE Expert Group (15 min). Finally, we will define next steps and activities and conclude with a wrap up of the whole Workshop (10 min) (Table 1).

Table 1. The agenda of the AIED 2025 Workshop "Ethical AI and Education: The need for international regulation to foster human rights, democracy and equity"

Sessions	Agenda Item	Duration (minutes)
1. Start	Welcome to the workshop	5'
	Introduction to the CoE work on AIED	10'
	Discussion of four questions (Q1 to Q4)	40'
	Summarising participants' input and wrap up	20'
2. World Café	Introduction to the World Café	5'
	First round of group discussions	25'
	Second round of group discussions	25'
	Third round of group discussions	25'
3. Plenary	Results from the group tables	30'
4. Closing	Collecting participants' input for future CoE work	15'
	Envisioning future steps and directions	10'

3.2 Target Audience

The workshop is open to all members of the AIED community, regardless of their level of experience. It welcomes both newcomers who may be exploring these issues for the first time and experienced experts who are already engaged in the field. By using a hybrid format that includes both in-person and online participation, the workshop aims to create space for inclusive and diverse discussions across the global AIED community.

4 Expected Outcomes and Next Steps

The workshop will collect and discuss key legal and organisational ideas that can help shape future strategies and guidelines for ethical and trustworthy AI. This includes identifying what is needed to ensure that AI use supports democratic values, both in education and in other areas. Participants will work together to explore what kinds of rules and frameworks are required and how they might be applied in practice.

The results of these discussions will be shared with the Council of Europe's expert group on AI and Education. They will help inform the group's final report and the legal texts that will be considered by the Council of Ministers at the end of 2025. Furthermore, all results will be published on the workshop website [3].

Additionally, there are plans to develop a joint publication based on the outcomes of the workshop. This will document the ideas and recommendations that emerge and make them available to a wider audience.

5 Conclusion

The unchecked use of AI in education threatens to undermine key societal values. As education plays a unique role in shaping democratic culture, it must be protected through deliberate, transparent, and enforceable regulation. This workshop seeks to move beyond abstract discussions and contribute concrete input to international governance processes.

We hope that the exchanges, collaboration and results of the workshop will strengthen the public awareness and required AIED regulation to facilitate responsible, fair and ethical AI use in education as well as they will foster AI literacy in all sectors and disciplines to enable education about AI and AI competence building at large-scale.

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

References

1. AIED 2023 Workshop Website: http://aied2023.learning-innovations.eu. Accessed 06 May 2025
2. AIED 2024 Workshop Website: http://aied2024.learning-innovations.eu. Accessed 06 May 2025

3. AIED 2025 Workshop Website: http://aied2025.learning-innovations.eu. Accessed 06 May 2025
4. Bond, M., et al.: A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. Int. J. Educ. Technol. Higher Educ. **21**(4) (2024). https://doi.org/10.1186/s41239-023-00436-z
5. Bozkurt, A., et al.: The Manifesto for Teaching and Learning in a Time of Generative AI: a critical collective stance to better navigate the future. Open Praxis **16**(4), 487–513 (2024). https://doi.org/10.55982/openpraxis.16.4.777
6. Bozkurt, A., et al.: Speculative Futures on ChatGPT and Generative Artificial Intelligence (AI): a collective reflection from the educational landscape. Asian J. Distance Educ. **18**(1), 53–130 (2023). https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/709/394
7. Council of Europe: Regulating Artificial Intelligence in education (2023a). https://rm.coe.int/regulating-artificial-intelligence-in-education-26th-session-council-o/1680ac9b7c
8. Council of Europe: The Transformative Power of Education: Universal Values and Civic Renewal. Resolutions of the 26th Session of the Council of Europe Standing Conference of Ministers of Education (28–29 September 2023). MED-26(2023)06 final (2023b). https://rm.coe.int/resolutions-26th-session-council-of-europe-standing-conference-of-mini/1680abee7f
9. CoE Report on Working Conference 2023. https://www.coe.int/en/web/education/-/working-conference-on-regulating-the-use-of-ai-systems-in-education. Accessed 06 May 2025
10. CoE Report on Working Conference 2024. https://rm.coe.int/artificial-intelligence-and-education-2nd-working-conference-provision/1680b314a3. Accessed 06 May 2025
11. Holmes, W., Persson, J., Chounta, I.-A., Wasson, B., Dimitrova, V.: Artificial intelligence and education: A critical view through the lens of human rights, democracy and the rule of law. Council of Europe (2022). https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd
12. Holmes, W., et al.: AI and Education. A view through the lens of human rights, democracy and the rule of law. Legal and organizational requirements. In: Artificial Intelligence in Education. Communications in Computer and Information Science, vol. 1831, pp. 79–84 (2023). https://doi.org/10.1007/978-3-031-36336-8_12
13. W Holmes I Tuomi 2022 State of the art and practice in AI in education Eur. J. Educ. 57 542 570 https://doi.org/10.1111/ejed.12533
14. A Nguyen HN Ngo Y Hong B Dang B-PT Nguyen 2023 Ethical principles for artificial intelligence in education Educ. Inf. Technol. 28 4221 4241 https://doi.org/10.1007/s10639-022-11316-w
15. Stracke, C.M.: Artificial intelligence and education: ethical questions and guidelines for their relations based on human rights, democracy and the rule of law. In: Burgos, D., et al. (Eds.), Radical Solutions for Artificial Intelligence and Digital Transformation in Education. Lecture Notes in Educational Technology, pp. 97–107 (2024). https://doi.org/10.1007/978-981-97-8638-1_7
16. Stracke, C.M., et al.: What is Artificial Intelligence (AI)? How can I use AI ethically at university? Edited by the German Network "Ethical Use of AI" (2024a). https://doi.org/10.5281/zenodo.10995669
17. CM Stracke I-A Chounta V Dimitrova B Havinga W Holmes 2024 Ethical AI and Education: the need for international regulation to foster human rights, democracy and the rule of law Artif. Intell. Educ. Commun. Comput. Inf. Sci. 2151 439 445 https://doi.org/10.1007/978-3-031-64312-5_55
18. CM Stracke I-A Chounta W Holmes 2024 Global trends in scientific debates on trustworthy and ethical artificial intelligence and education Artif. Intell. Educ. Commun. Comput. Inf. Sci. 2150 254 262 https://doi.org/10.1007/978-3-031-64315-6_21

19. CM Stracke 2025 Analysis of artificial intelligence policies for higher education in Europe Int. J. Interact. Multimedia Artif. Intell. 9 2 124 137 https://doi.org/10.9781/ijimai.2025.02.011
20. Tlili, A., et al.: Taming the monster: how can open education promote the effective and safe use of generative AI in education? J. Learn. Dev. **11**(3), 398–413 (2024). https://doi.org/10.56059/jl4d.v11i3.1657
21. Xiao, J., et al.: Venturing into the unknown: critical insights into grey areas and pioneering future directions in educational generative AI research. TechTrends (2025). https://doi.org/10.1007/s11528-025-01060-6

The Second International Workshop on Multimodal Artificial Intelligence in Education (MAIEd'25)

Daniele Di Mitri[1(✉)], Namrata Srivastava[2,3], Gloria Fernandez Nieto[3], Vanessa Echeverria[3,4], Ruth Cobos[5], Ashwin Tudur Sadashiva[2], Daniel Spikol[6], Kester Yew Chong Wong[7], Qi Zhou[7], and Mutlu Cukurova[7]

[1] German University of Digital Science, Potsdam, Germany
daniele.dimitri@german-uds.de
[2] Vanderbilt University, Nashville, USA
[3] Monash University, Melbourne, Australia
[4] Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador
[5] Universidad Autónoma de Madrid, Madrid, Spain
[6] University of Copenhagen, Copenhagen, Denmark
[7] University College London, London, UK

Abstract. Following its launch in 2021, this workshop seeks to gather fresh insights into employing Artificial Intelligence (AI) systems and intelligent tutors for education and learning through the use of multimodal data sources. The Multimodal Artificial Intelligence in Education (MAIEd) workshop builds upon the CrossMMLA workshop series at the Learning Analytics & Knowledge Conference and the Multimodal Immersive Learning Systems (MILeS) at the European Conference of Technology Enhanced Learning. The MAIEd workshop calls for new empirical studies, even if they are in their early stages of development. It welcomes novel experimental designs, theoretical contributions, and practical demonstrations that can prove the use of multimodal and multisensor devices "beyond mouse and keyboard" in learning contexts for automatic feedback generation, adaptation, and personalisation in learning. Through a call for proposals, we seek to engage the scientific community in opening up the scope of AI in Education towards novel and diverse data sources.

Keywords: Multimodality · Automatic Feedback · Sensor-based systems · Educational scenarios

1 Background

Learning is a complex, multidimensional process involving various modalities, including visual, auditory, kinesthetic, and social elements. This multifaceted nature of learning means that individuals often benefit from a combination of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 292–299, 2025.
https://doi.org/10.1007/978-3-031-99267-4_38

methods, such as reading, listening, hands-on activities, and collaborative discussions, to fully grasp and retain new information.

Formal education often relies on traditional methods of instruction and fails to recognize the crucial role of multimodality in the learning process.

In the AIEd research community, multimodality refers to integrating multiple data sources—such as text, speech, facial expressions, gestures, and sensor data to build richer models of learner behavior and engagement. Traditionally, many AIEd approaches, such as Intelligent Tutoring Systems (ITS), have relied on unimodal data, like textual inputs or clickstream logs, which capture only a limited view of the learning experience. In contrast, multimodal approaches can provide a more holistic picture by combining diverse data types, enabling systems to detect subtle cues (for example, affective states through facial expressions or tone of voice) that enhance personalization and adaptivity in educational environments.

While multimodal AI has been widely explored in K-12 and higher education, its role in adult learning environments, such as simulation-based training for healthcare, AI-driven professional training, and workplace learning settings, remains underexplored. These settings require multimodal AI to facilitate real-time feedback, hands-on skill development, and personalized learning pathways in high-stakes environments.

Expanding the scope of multimodal AI, we aim to advance the scientific discussion by drawing special attention to AI techniques for analyzing multimodal data in diverse learning contexts. In addition to traditional classroom-based applications, this includes simulation-based training, professional skill development, and workplace learning environments, where multimodal AI plays a crucial role in real-time decision-making and adaptive feedback.

This area remains relatively niche for several reasons. The technical challenges of aligning and processing heterogeneous data streams—each with its own sampling rates, noise characteristics, and context-specific meanings—are nontrivial. In multimodal AI systems, integrating diverse modalities like speech, vision, motion data (from embodied learning environments), and behavioral interactions introduces challenges in data alignment, fusion, and translation. These challenges are particularly pronounced in collaborative and narrative-centered learning environments, where multimodal AI must track interactions across multiple learners, evolving scenarios, and diverse perspectives in real time.

These complexities impact model robustness, real-time adaptability, and the meaningful interpretation of learner states. Addressing them requires novel techniques for synchronization, cross-modal learning, and data fusion strategies to enable seamless inference from heterogeneous data sources. While promising work has emerged in multimodal learning analytics, its integration into fully operational AIEd systems remains slow, partly due to the difficulty of merging insights from multiple modalities into coherent, actionable models.

Recent advancements in multisensor devices and multimodal data provide a more detailed digital representation of the learner [6]. This shift has sparked increasing interest in the educational research community, particularly in multi-

modal and multi-sensor interaction methods. Physiological sensors, for instance, can offer valuable insights that contextualize learning performance and strategies [12]

The AIEd community has recently increasingly utilized pre-trained generative models, such as large language models (LLMs). Compared to other intelligent methods such as ITS, LLMs are much more flexible in how they are trained, how they can be used, and how the final user can fine-tune them through prompting. In particular, multimodal LLMs (MLLMs) and multimodal large vision models (MLVMs) have become more popular. These models can handle multimodal input beyond simple text prompts, including images, videos, and sound. The MLLM models leverage the multimodality concept, inspired by how humans communicate and interact using multiple communication channels.

AI tutors based on generative AI enable students to receive automated feedback in a more personalized and efficient manner. This technology primarily utilizes text as the interaction modality, allowing for immediate responses to students' inquiries and assignments.

GenAI can process multimodal data, ranging from text to images and audio, providing richer insights into learning interactions across diverse educational settings [11]. For example, GenAI can automate content generation, enabling educators to quickly develop personalized, multimodal learning materials and enhance student engagement by offering real-time feedback based on behavioural and performance data [11]. However, these advancements necessitate deeper reflection to address the ethical, technical, and societal issues arising from using GenAI models [2].

With the use of multimodal data and the rise of GenAI techniques in AIEd, ethical considerations become paramount. Data privacy and security are significant concerns when dealing with sensitive data such as video, voice, and biometric information. These models' cost, scalability, and adaptability across different educational contexts also require further inquiry. Bias in AI models is another key challenge, as GenAI could unintentionally perpetuate inequities in learning environments if not managed carefully [2,3].

This workshop will bring together researchers and practitioners to discuss how multimodal data can be effectively and responsibly integrated into AIED. Discussions will focus on designing ethical, transparent, and scalable AIED tools while addressing privacy and bias concerns. Integrating GenAI with human expertise, ensuring open learner models for transparency, and employing hybrid AI-human systems could mitigate many of these issues [15].

The MAIEd workshop series [5], together with the CROSSMMLA workshop series [1,8,10,16], has been pivotal in advancing MMLA research, and this year's workshop will build on this legacy by evaluating how GenAI is shaping the future of MMLA, offering practical insights into enhancing learning analytics while maintaining ethical standards and promoting equitable access to AI technologies in education.

In this workshop, acknowledging the previous body of research, we do not aim to take a side between one or the other, but we are open to any relevant

scientific finding from either side. We want to advance the scientific discussion about multimodality for learning by drawing special attention to AI techniques for analysing multimodal data for learning.

2 Related Research

To illustrate the progress of the field to be covered in the workshop, we include next some examples of multimodal educational systems that will be introduced in the workshop to start the discussion:

- Presentable, which is a new web-based system for presentation training
- The CPR Tutor, which provides real-time audio feedback during cardiopulmonary resuscitation training [7]
- The KUMITRON system [9] uses physiological signals and computer vision to monitor the performance in a Karate combat; this information can be used to give audio feedback during training that complements visual information in real-time.
- The Calligraphy Trainer [13] a tablet application to teach people how to write in a foreign alphabet using a capacitive pen and electromyographic information
- The Table Tennis Tutor [14], which captures motion data and tracks the skeletal points of the body to observe the execution of fundamental table tennis strokes, can also be used to provide automatic feedback during training.
- GePeTo (Generative AI based Personalized Guidance Tool): this system provides personalized, actionable feedback tailored to individual student performance in both online and face-to-face learning tasks [4].
- Collaborative Problem Solving (CPS) Diagnostic, which uses a transformer-based model to detect CPS indicators from multimodal audio data.
- A collaborative learning analytics system, which detects effective collaborative learning interactions through video and audio data in a real-world face-to-face collaborative learning higher education context. [17]

3 Workshop

3.1 Scope & Topics

At the MAIEd workshop, we want to discuss which scientific, state-of-the-art ideas and approaches are being pursued and which impacts we expect on educational technologies and education. We are especially interested in contributions targeting the intersection of these two fields of AI and multimodal interaction. We are looking for original contributions that advance the state of the art in theories, technologies, methods, and knowledge towards developing multimodal intelligent tutors. Concrete topics of interest for this year's MAIEd workshop include but are not limited to:

- Multimodal AI Architectures and Systems

  - Multimodal Intelligent Tutoring Systems or User Interfaces
  - AI Multimodal Architectures and Scalable Strategies
  - Multimodal Learner Modeling and Affective Computing
  - Adaptive Feedback, Guidance, and Process in Multimodal Learning
  - Artificial Intelligence for Learning Analytics
  - AI in Professional and Workplace Learning
- Multimodal Learning Environments
  - Multimodality in Augmented, Virtual, and Mixed Reality
  - Multimodal AI for Embodied Learning Environments
  - Multimodal AI in Narrative-Centered Learning Environments
  - Collaborative Learning and AI-Driven Group Interaction Analysis
- Methods and Technical Challenges
  - Big Data-driven Visual Analytics for Learning
  - Multimodal Data Fusion and Cross-Modal Learning in AI Tutors
  - Error Detection and Classification for Multimodal Data
  - Cognitive Load in Multimodal Interaction with Intelligent Tutoring Systems
  - Explainability, Trust, and Safety in Multimodal Intelligent Tutoring Systems
  - Bias, Fairness, and Ethical AI in Multimodal Learning

In addition, and in line with the AIED 2025 conference theme, we aim to critically examine and reflect on issues of equity, diversity, and inclusion in MAIEd. This includes exploring how MAIEd technologies can be designed and implemented to support learners from diverse backgrounds, address biases in AI-driven educational tools, and promote fair and accessible learning opportunities for all. By engaging with these important considerations, we seek to ensure that MAIEd contributes to a more inclusive and equitable educational landscape.

3.2 Organisational Details

The workshop will be conducted in person and as a "full day" in three parts. If allowed, the workshop will also offer a hybrid option. The submitted contributions will be presented in the morning in the first part. In the second part, there will be an interactive discussion, using either smaller group discussion. In the last part, after a lunch break, there will be a keynote speaker followed by an open discussion. This workshop aims to publish the accepted contributions through the CEUR proceedings. For this reason, we invite the authors to follow the CEUR's proceeding guidelines when submitting their contributions.

4 Organisers and Program Committee

Prof. Dr. Daniele Di Mitri is a professor of Multimodal Learning Technologies and leader of the "Augmented Feedback" research lab at the German University of Digital Science. He is an associate researcher at the DIPF - Leibniz Institute for Research and Information in Education and a lecturer at the Goethe

University of Frankfurt, Germany. His current research focuses on developing AI-driven, multimodal learning technologies to enhance digital education.

Dr. Namrata Srivastava Dr. Namrata Srivastava is currently a Research Scientist at the Vanderbilt University. Prior to joining this role, she honed her research skills as a Postdoctoral Researcher in the Penn Center for Learning Analytics (PCLA) lab at the University of Pennsylvania, United States, and also as a Research Fellow in the Centre for Learning Analytics at Monash University (CoLAM) in Australia for three years.

Dr. Gloria Fernández Nieto is a Research Fellow at Monash University, Australia. Her research focuses on providing embodied learning feedback through data storytelling. Her interests include human-centred AI and leveraging Generative AI to support teachers in knowledge management tasks.

Dr. Vanessa Echeverria is a Lecturer at RMIT University and a Research Associate at Monash University. Her research interests are in human-centred AI, multimodal learning analytics for feedback, and collaboration analytics in physical spaces.

Dr. Ruth Cobos is an Associate Professor in the Computer Engineering Department at the Universidad Autónoma de Madrid (UAM, Spain). She is a member of the Spanish Network of Learning Analytics (SNOLA) and the Group for Advanced Interactive Tools (GHIA) at UAM, focusing on research areas related to the development of AI-driven, multimodal learning technologies to support both online and face-to-face learners in their learning tasks.

Dr. Ashwin Tudur Sadashiva Dr. Ashwin Tudur Sadashiva is a Research Scientist at Vanderbilt University's Institute for Software Integrated Systems, Learning Incubator: a Vanderbilt Endeavor (LIVE) Lab, and Open-Ended Learning Environments Lab. His research focuses on AI-driven educational systems that integrate multimodal data sources, including computer vision, speech, and behavioral interactions, to enhance learning analytics, intelligent tutoring, and adaptive feedback. He works on bias mitigation in AI-driven learning, ethical AI in education, and AI literacy.

Dr. Daniel Spikol is an Associate Professor at the Center of Digital Education at the University of Copenhagen. Spikol's position is divided between the Departments of Computer Science and Science Education. His current research investigates and develops tools for collaboration for learning with multimodal learning analytics and Artificial Intelligence. His work develops technologies that support learning, play, and reflection through computational tinkering and thinking.

Prof. Dr. Mutlu Cukurova is a Professor of Learning and Artificial Intelligence at University College London. He directs the (UCL Analytics & AI for Learning) UCLAIT team and engages in policy-making activities as an external expert (including UNESCO, IAEA, and the EC). He was the programme co-chair of the International Conference of AI in Education in 2020, is the Editor of the British Journal of Educational Technology, and an Associate Editor of the International Journal of Child-Computer Interaction.

Dr. Qi Zhou is a postdoctoral research fellow and part of the UCLAIT team at the UCL Knowledge Lab, University College London, UK. His research focuses

on applying multimodal learning analytics to support collaborative learning. He is also interested in investigating human-AI complementarity in supporting both learning and teaching.

Kester Yew Chong Wong is a PhD student and part of the UCLAIT team at the UCL Knowledge Lab, University College London, UK. His research focuses on AI-based support for Collaborative Problem Solving (CPS). Kester is also a Teaching Assistant at the Nanyang Technological University, National Institute of Education (NTU/NIE) in Singapore under the Learning Sciences and Assessment Academic Group (LSA AG).

References

1. Proceedings of the Second Multimodal Learning Analytics Across (Physical and Digital) Spaces co-located with the 8th International Learning Analytics and Knowledge Conference, CrossMMLA@LAK 2018, Sydney, Australia, March 06, 2018. CEUR Workshop Proceedings, vol. 2163. CEUR-WS.org (2018). http://ceur-ws.org/Vol-2163
2. Alwahaby, H., Cukurova, M., Papamitsiou, Z., Giannakos, M.: The evidence of impact and ethical considerations of multimodal learning analytics: a systematic literature review. In: Giannakos, M., Spikol, D., Di Mitri, D., Sharma, K., Ochoa, X., Hammad, R. (eds.) The Multimodal Learning Analytics Handbook, pp. 289–325. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-08076-0_12
3. Ashwin, T., Biswas, G.: Identifying and mitigating algorithmic bias in student emotional analysis. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds) International Conference on Artificial Intelligence in Education, pp. 89–103. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64299-9_7
4. Becerra, A., Mohseni, Z., Sanz, J., Cobos, R.: A generative AI-based personalized guidance tool for enhancing the feedback to MOOC learners. In: 2024 IEEE Global Engineering Education Conference (EDUCON), pp. 1–8 (2024). https://doi.org/10.1109/EDUCON60312.2024.10578809, https://ieeexplore.ieee.org/document/10578809, iSSN: 2165-9567
5. Di Mitri, D., et al.: Proceedings of the First International Workshop on Multimodal Artificial Intelligence in Education (MAIED 2021), vol. 2902, p. 112. CEUR Workshop Proceedings, Aachen, Germany (2021). http://ceur-ws.org/Vol-2902/
6. Di Mitri, D., Schneider, J., Specht, M., Drachsler, H.: From signals to knowledge: a conceptual model for multimodal learning analytics. J. Comput. Assisted Learn. **34**(4), 338–349 (2018). https://doi.org/10.1111/jcal.12288, https://onlinelibrary.wiley.com/doi/10.1111/jcal.12288
7. Di Mitri, D., Schneider, J., Trebing, K., Sopka, S., Specht, M., Drachsler, H.: Real-time multimodal feedback with the CPR tutor. In: Bittencourt, I., Cukurova, M., Muldner, K. (eds.) Artificial Intelligence in Education (AIED'2020), pp. 141–152. Springer, Cham, Switzerland (2020). https://doi.org/10.1007/978-3-030-52237-7_12, http://link.springer.com/10.1007/978-3-030-52237-7_12
8. Di Mitri, D.D., Srivastava, N., Martínez-Maldonado, R., Cukurova, M., Spikol, D. (eds.): Proceedings of the 6th Workshop on Leveraging Multimodal Data for Generating Meaningful Feedback (CROSSMMLA 2023) at the 13th International

Learning Analytics & Knowledge (LAK 2023), Arlington, Texas, USA, March 13-17, 2023, CEUR Workshop Proceedings, vol. 3439. CEUR-WS.org (2023). https://ceur-ws.org/Vol-3439

9. Echeverria, J., Santos, O.C.: Toward modeling psychomotor performance in karate combats using computer vision pose estimation. Sensors **21**(24), 8378 (2021). https://doi.org/10.3390/s21248378, https://www.mdpi.com/1424-8220/21/24/8378, number: 24 Publisher: Multidisciplinary Digital Publishing Institute
10. Giannakos, M., Spikol, D., Molenaar, I., Di Mitri, D., Sharma, K., Ochoa, X., Hammad, R.: Preface: CrossMMLA in practice: collecting, annotating and analyzing multimodal data across spaces. In: CEUR Workshop Proceedings, vol. 2610. CEUR Workshop Proceedings (2020). http://ceur-ws.org/Vol-2610/xpreface.pdf
11. Giannakos, M., et al.: The promise and challenges of generative AI in education. Behav. Inf. Technol., 1–27 (2024). https://doi.org/10.1080/0144929X.2024.2394886, https://doi.org/10.1080/0144929X.2024.2394886, publisher: Taylor & Francis eprint: https://doi.org/10.1080/0144929X.2024.2394886
12. Giannakos, M.N., Sharma, K., Pappas, I.O., Kostakos, V., Velloso, E.: Multimodal data as a means to understand the learning experience. Int. J. Inf. Manage. **48**(February), 108–119 (2019). https://doi.org/10.1016/j.ijinfomgt.2019.02.003, https://doi.org/10.1016/j.ijinfomgt.2019.02.003, publisher: Elsevier
13. Limbu, B.H., Jarodzka, H., Klemke, R., Specht, M.: Can you ink while you blink? Assessing mental effort in a sensor-based calligraphy trainer. Sensors (Basel, Switzerland) **19**(14), E3244 (2019). https://doi.org/10.3390/s19143244
14. Mat Sanusi, K.A., Di Mitri, D., Limbu, B., Klemke, R.: Table tennis tutor: forehand strokes classification based on multimodal data and neural networks. Sensors **21**(9), 3121 (2021). https://doi.org/10.3390/s21093121, https://www.mdpi.com/1424-8220/21/9/3121, number: 9 Publisher: Multidisciplinary Digital Publishing Institute
15. Prinsloo, P., Slade, S., Khalil, M.: Multimodal learning analytics—In-between student privacy and encroachment: a systematic review. Br. J. Educ. Technol. **54**(6), 1566–1586 (2023). https://doi.org/10.1111/bjet.13373, https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13373
16. Spikol, D., et al.: CROSSMMLA Futures: collecting and analysing multimodal data across the physical and the virtual. In: 11th International Learning Analytics and Knowledge Conference (LAK'21) (2021). https://www.solaresearch.org/wp-content/uploads/2021/04/LAK21_CompanionProceedings.pdf, vol. 2021. https://discovery.ucl.ac.uk/id/eprint/10126482/
17. Zhou, Q., Suraworachet, W., Cukurova, M.: Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. Educ. Inf. Technol. **29**(1), 1071–1098 (2024). https://doi.org/10.1007/s10639-023-12315-1, https://link.springer.com/10.1007/s10639-023-12315-1



NLP and Generative AI for Language Learning and Assessment: Synergies Between Research and Practice

Zheng Yuan[1(✉)], Mariano Felice[2], Yunshi Lan[3], and Qiao Wang[4]

[1] The University of Sheffield, Sheffield, UK
zheng.yuan1@sheffield.ac.uk
[2] British Council, London, UK
Mariano.Felice@britishcouncil.org
[3] East China Normal University, Shanghai, China
yslan@dase.ecnu.edu.cn
[4] Hosei University, Tokyo, Japan
judy.wang@hosei.ac.jp

Abstract. This tutorial explores the evolving intersection of Natural Language Processing (NLP), Generative AI, and language education. It provides a comprehensive overview of how recent advances in NLP – from automated item generation and grammatical error correction to multimodal tutoring systems and real-time feedback – are transforming language learning and assessment. A central focus is on the potential of Large Language Models (LLMs) for personalised instruction, interactive evaluation, and inclusive learning experiences. Drawing from both academic research and real-world applications, the tutorial will cover cutting-edge methodologies, system design, and ethical considerations, offering participants practical insights into the responsible development and deployment of AI-driven educational tools. Through case studies and demonstrations, attendees will engage with emerging technologies that support more effective, equitable, and adaptive language learning.

Keywords: Educational NLP · Generative AI · Language Learning · Language Assessment · Automated Feedback · Ethical AI in Education

1 Tutorial Chairs

Zheng Yuan is an Associate Professor in Natural Language Processing (NLP) at the University of Sheffield. Before joining Sheffield, Zheng was an Assistant Professor at King's College London and a Research Associate at the University of Cambridge. Her research focuses on machine learning, NLP, and their use in educational applications, including automated assessment, grammatical error correction, and feedback generation. Zheng serves as the Treasurer of the Association for Computational Linguistics (ACL) Special Interest Group for Building Educational Applications (SIGEDU) and has been involved in co-organising the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 300–307, 2025.
https://doi.org/10.1007/978-3-031-99267-4_39

yearly Workshop on Innovative Use of NLP for Building Educational Applications (BEA) and various shared tasks including the BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues[1].

Mariano Felice is a Senior Researcher and Data Scientist for Language Assessment and Learning at the British Council. His work involves looking at how AI and NLP can be used to improve computer-assisted learning and testing, from curating datasets and building models to supporting colleagues in the adoption of new technology. Mariano is also a Visiting Researcher at the universities of Reading and Cambridge and was a Research Associate in the ALTA Institute before joining the ELR group at the British Council. He has a PhD in computer science from the University of Cambridge and has worked extensively on NLP for language learning and assessment, including grammatical error correction, automatic error typing, system evaluation, automated cloze test generation and item calibration. Mariano has published many scientific papers in top-tier NLP conferences, given talks and tutorials in different venues around the world and is a regular reviewer for workshops, journals and conferences in his field. Mariano was one of the organisers of the BEA-2019 Shared Task on Grammatical Error Correction[2].

Yunshi Lan is an Associate Professor of School of Data Science and Engineering at East China Normal University. She obtained her Ph.D. degree from the School of Computing and Information Systems at Singapore Management University. Her research interests lie in NLP with a focus on educational NLP, question answering, and large language models. She has published a number of papers in international conferences and journals such as ACL, AAAI, IJCAI, ACM MM, NeurIPS, and IEEE TKDE. She served as an Area Chair, a Session Chair and a PC member for a number of journals and conferences, including AAAI, IJCAI, ACL, EMNLP, etc.

Qiao Wang is an Associate Professor at the School of Economics, Hosei University in Japan. Before joining Hosei, she worked at the Center for English Language Education, Faculty of Science and Engineering at Waseda University. She obtained her Ph.D. from Kyoto University, focusing on Computer-Assisted Language Learning (CALL). She is a practitioner of academic English and her research is focused on the utilization of technology in English language classrooms. Her endeavours in recent years include automated vocabulary quiz generation, automated writing evaluation systems, grammar diagnostic testing, grammar error detection and correction and game-based language learning. She has developed various relevant web applications to facilitate both students and teachers in language classrooms. She has also published in prestigious journals and NLP conference proceedings.

[1] https://sig-edu.org/sharedtask/2023.
[2] https://www.cl.cam.ac.uk/research/nl/bea2019st/.

2 Tutorial Content and Themes

This tutorial provides a comprehensive overview of how NLP is transforming language learning and assessment. It will explore past developments, current state-of-the-art methodologies, and future directions in the field. The tutorial will cover multi-modal and knowledge-enhanced tutoring systems, automatic item generation, grammatical error correction, automated scoring and feedback generation, and ethical considerations in educational NLP.

A key focus will be on Generative AI and LLMs, which have brought transformative changes to educational technology. The tutorial will discuss how these models can be adapted for personalised learning, interactive assessments, and real-time feedback mechanisms. Topics such as prompt engineering, few-shot learning, and instruction-tuned models will be explored to highlight their potential for adaptive and equitable education.

Through discussions, case studies, and hands-on demonstrations, participants will gain insights into how NLP-powered tools are reshaping education, ensuring personalisation, inclusivity, and accessibility.

3 Significance and Relevance to the AIED Community

With AI playing a pivotal role in education, NLP is crucial for creating intelligent systems that enhance both teaching and learning. This tutorial aligns with the AIED 2025 theme of AI as a Catalyst for Inclusive, Personalised, and Ethical Education: Empowering Teachers and Students for an Equitable Future. Key contributions to the AIED community include addressing real-world challenges in language learning and assessment, demonstrating state-of-the-art AI-driven language education tools, and providing insights into the ethical deployment of NLP in education.

The tutorial will be beneficial to AI researchers, educators, technologists, and policymakers interested in leveraging NLP for equitable and effective learning.

4 Tutorial Format and Activities

1. Introduction to Educational NLP – 20 mins
   (a) The role of NLP in education: teaching and learning & assessment
   (b) Overview of typical NLP tasks for language education
2. Teaching and learning – 70 mins
   (a) Tutoring systems
      i Multi-modal QA systems [26,29]
      ii Knowledge-enhanced QA systems [21,25,37]
      iii LLM-based dialogue systems [11,30,43]
      iv Challenges and future directions [27]
   (b) Automatic item generation
      i Cloze items and multiple choice questions [18,46,47]
      ii Open-ended questions [2,7,40]

   iii Item difficulty prediction [17,50,51]
   iv Challenges and future directions [27]

3. Assessment – 70 mins
   (a) Grammatical error correction
      i Sequence-to-sequence models [20,53,54]
      ii Sequence-to-edit models [32,35]
      iii LLMs-based approaches [12,39]
      iv Challenges and future directions [5,48]
   (b) Automated scoring
      i Automated essay scoring [19,22,31,45]
      ii Automated scoring of speech [3,23,55]
      iii Feedback generation [13,38,52]
      iv Challenges and future directions [27]

4. Ethics – 20 mins
   (a) Ethical concerns
   (b) Legislation and recommendations [9,10,34,41]
   (c) Ethics by design [4,8,24]
   (d) Ethical AI and assessment organisations [6,14,16]
5. Moving forward – 30 mins
   (a) Adaptive NLP techniques for education [33,42]
   (b) Interpretable educational NLP [1,15,49]
   (c) Game-based language learning [28,36,44]

5 Expected Target Audience and Anticipated Attendance

This tutorial is designed for AI and NLP researchers interested in educational applications, educators and instructional designers looking to integrate AI in classrooms, EdTech professionals developing AI-driven language learning tools, and graduate students and early-career researchers exploring NLP for education. Given the rapid growth of this field, we anticipate over 50 participants.

6 Funding Solicitation Plans

No external funding is required for the tutorial. However, the organisers will explore potential sponsorships for open-access materials and travel grants for underrepresented groups.

7 Plans for Advertising the Event

The tutorial will be advertised through the AIED 2025 conference website and social media platforms, SIGEDU, and other NLP/AI mailing lists, institutional websites and networks of the tutorial organisers, and through promotion at pre-AIED conferences and workshops.

8 Plans for Tutorial Dissemination

Tutorial materials, including slides and resources, will be publicly available via the organisers' websites and tutorial GitHub repositories[3].

9 Ethical Concerns

The content presented in this tutorial is based on publicly available research, much of which has already undergone ethical review during the publication process. Furthermore, none of the materials generated for this tutorial require the involvement of human participants or pose any risks to the environment. As a result, no additional ethical approval is necessary.

Beyond compliance with ethical guidelines, this tutorial will actively engage with critical ethical concerns surrounding the application of NLP in education. Key issues such as data privacy, algorithmic bias, and the digital divide will be addressed in depth. We will explore the implications of these challenges and provide practical guidelines for mitigating risks, ensuring that NLP-driven educational technologies are developed and deployed in a responsible and equitable manner.

By integrating ethical considerations into our discussions, we aim to promote transparency, inclusivity, and fairness in educational NLP. Our goal is to equip researchers and practitioners with the necessary knowledge to create solutions that benefit all students and teachers while minimising potential harm.

References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. In: Proceedings of NAACL (2016)
2. Attali, Y., et al.: The interactive reading task: transformer-based automatic item generation. Front. Artif. Intell. (2022)
3. Bamdev, P., Grover, M.S., Singla, Y.K., Vafaee, P., Hama, M., Shah, R.R.: Automated speech scoring system under the lens. Int. J. Artif. Intell. Educ. (2023)
4. Brey, P., Dainow, B.: Ethics by design for artificial intelligence. AI and Ethics (2023)
5. Bryant, C., Yuan, Z., Qorib, M.R., Cao, H., Ng, H.T., Briscoe, T.: Grammatical error correction: a survey of the state of the art. Comput. Linguist. (2023)
6. Burstein, J.: The Duolingo English test responsible AI standards (2023). https://go.duolingo.com/ResponsibleAI. [Updated March 29, 2024]
7. Chakrabarty, T., Lewis, J., Muresan, S.: CONSISTENT: open-ended question generation from news articles. In: Findings of EMNLP (2022)
8. Commission, E.: Ethics by design and ethics of use approaches for artificial intelligence (2021)
9. Commission, E.: Eu artificial intelligence act (2024)

[3] https://aied2025-lla-tutorial.github.io/.

10. Commission, E., Directorate-General for Education, Youth, S., Culture: ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators. Publications Office of the European Union (2022)
11. Dan, Y., et al.: EduChat: a large-scale language model-based chatbot system for intelligent education. arXiv preprint arXiv:2308.02773 (2023)
12. Davis, C., et al.: Prompting open-source and commercial language models for grammatical error correction of English learner text. arXiv preprint arXiv:2401.07702 (2024)
13. Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., De Weerdt, J.: A review of automated feedback systems for learners: classification framework, challenges and opportunities. Comput. Educ. (2021)
14. Dieterle, E., Dede, C., Walker, M.: The cyclical ethical effects of using artificial intelligence in education. AI Soc. (2024). https://doi.org/10.1007/s00146-022-01497-w
15. Du, H., et al.: FlaCGEC: a Chinese grammatical error correction dataset with fine-grained linguistic annotation. In: Proceedings of CIKM (2023)
16. Felice, M.: Human-Centred AI: lessons for English learning and assessment (2024). https://qrco.de/bfmW8S, British Council
17. Felice, M., Karaoz, Z.D.: The British council submission to the BEA 2024 shared task. In: Proceedings of BEA (2024)
18. Felice, M., Taslimipoor, S., Buttery, P.: Constructing open cloze tests using generation and discrimination capabilities of transformers. In: Findings of ACL (2022)
19. Ifenthaler, D.: Automated Essay Scoring Systems (2022)
20. Junczys-Dowmunt, M., Grundkiewicz, R., Guha, S., Heafield, K.: Approaching neural grammatical error correction as a low-resource machine translation task. In: Proceedings of NAACL (2018)
21. Kafle, K., Price, B., Cohen, S., Kanan, C.: DVQA: understanding data visualizations via question answering. In: Proceedings of CVPR (2018)
22. Ke, Z., Ng, V.: Automated essay scoring: a survey of the state of the art. In: Proceedings of IJCAI (2019)
23. Khabbazbashi, N., Xu, J., Galaczi, E.D.: Opening the Black Box: exploring automated speaking evaluation (2021)
24. Kieslich, K., Keller, B., Starke, C.: Artificial intelligence ethics by design. evaluating public perception on the importance of ethical design principles of artificial intelligence. Big Data Soc. (2022)
25. Lan, Y., et al.: Complex knowledge base question answering: a survey. IEEE TKDE (2023)
26. Lan, Y., Li, X., Liu, X., Li, Y., Qin, W., Qian, W.: Improving zero-shot visual question answering via large language models with reasoning question prompts. In: Proceedings of ACM MM (2023)
27. Lan, Y., et al.: Survey of natural language processing for education: taxonomy, systematic review, and future trends. arXiv preprint arXiv:2401.07518 (2024)
28. Li, K., Peterson, M., Wang, Q., Wang, H.: Mapping the research trends of digital game-based language learning (DGBLL): a scientometrics review. Comput. Assisted Lang. Learn. (2023)
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
30. Lu, P., et al.: Learn to explain: multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)
31. Ludwig, S., Mayer, C., Hansen, C., Eilers, K., Brandt, S.: Automated essay scoring using transformer models. Psych (2021)

32. Mesham, S., Bryant, C., Rei, M., Yuan, Z.: An extended sequence tagging vocabulary for grammatical error correction. In: Findings of EACL (2023)
33. Naeiji, A.: Question generation using sequence-to-sequence model with semantic role labels. In: Proceedings of EACL (2022)
34. OECD: Recommendation of the council on artificial intelligence (2023)
35. Omelianchuk, K., Atrasevych, V., Chernodub, A., Skurzhanskyi, O.: GECToR – grammatical error correction: tag, not rewrite. In: Proceedings of BEA (2020)
36. Peterson, M., Wang, Q., Mirzaei, M.S.: The use of network-based virtual worlds in second language education: a research review. IGI Global (2023). https://doi.org/10.4018/978-1-6684-7597-3.ch011
37. Ram, O., Kirstain, Y., Berant, J., Globerson, A., Levy, O.: Few-shot question answering by pretraining span selection. In: Proceedings of ACL-IJCNLP (2021)
38. Shadiev, R., Feng, Y.: Using automated corrective feedback tools in language learning: a review study. Interact. Learning Environ. (2023)
39. Sottana, A., Liang, B., Zou, K., Yuan, Z.: Evaluation metrics in the era of GPT-4: reliably evaluating large language models on sequence to sequence tasks. In: Proceedings of EMNLP (2023)
40. Srivastava, M., Goodman, N.: Question generation for adaptive education. In: Proceedings of ACL-IJCNLP (2021)
41. UNESCO: Recommendation on the ethics of artificial intelligence (2021)
42. Uto, M., Tomikawa, Y., Suzuki, A.: Difficulty-controllable neural question generation for reading comprehension using item response theory. In: Proceedings of BEA (2023)
43. Wang, L., et al.: T-SciQ: teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In: Proceedings of AAAI (2024)
44. Wang, Q.: The role of classroom-situated game-based language learning in promoting students' communicative competence. Int. J. Comput.-Assist. Lang. Learn. Teach. **10**(2) (2020). https://doi.org/10.4018/ijcallt.2020040104
45. Wang, Q., Gayed, J.: Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. Comput. Assist. Lang. Learning (2024)
46. Wang, Q., Rose, R., Orita, N., Sugawara, A.: Automated generation of multiple-choice cloze questions for assessing English vocabulary using GPT-turbo 3.5. arXiv preprint arXiv:2403.02078 (2024)
47. Wang, Q., Rose, R., Sugawara, A., Orita, N.: Evaluation of automated vocabulary quiz generation with VocQGen. Vocabulary Learning Instruct. (2025)
48. Wang, Q., Yuan, Z.: Assessing the efficacy of grammar error correction: a human evaluation approach in the Japanese context. In: Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) (2024)
49. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of NeurIPS (2023)
50. 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 BEA (2024)
51. 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 BEA (2024)

52. Yoon, S.Y., Hsieh, C.N., Zechner, K., Mulholland, M., Wang, Y., Madnani, N.: Toward automated content feedback generation for non-native spontaneous speech. In: Proceedings of BEA (2019)
53. Yuan, Z., Briscoe, T.: Grammatical error correction using neural machine translation. In: Proceedings of NAACL (2016)
54. Yuan, Z., Taslimipoor, S., Davis, C., Bryant, C.: Multi-class grammatical error detection for correction: a tale of two systems. In: Proceedings of EMNLP (2021)
55. Zechner, K., Evanini, K.: Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech. Taylor & Francis (2019)

Fair4AIED 2025: First International Workshop on Fairness in Algorithmic Decision-Making for Education

Clara Belitz[1], Nigel Bosch[1], Lan Jiang[1], Ioana Jivet[2], Badmavasan Kirouchenassamy[3], HaeJin Lee[1], Barbara Leporini[4], Francesca Maridina Malloci[5], Mirko Marras[5], Daniela Rotelli[3(✉)], and Frank Stinar[1]

[1] University of Illinois Urbana–Champaign, Champaign, USA
{cbelitz2,pnb,lanj3,haejin2,fstinar2}@illinois.edu
[2] CATALPA, FernUniversität in Hagen, Hagen, Germany
ioana.jivet@fernuni-hagen.de
[3] Sorbonne University, Paris, France
{badmavasan.kirouchenassamy,daniela.rotelli}@lip6.fr
[4] University of Pisa & ISTI-CNR, Pisa, Italy
barbara.leporini@unipi.it
[5] University of Cagliari, Cagliari, Italy
{francescam.malloci,mirko.marras}@unica.it

Abstract. The applications of artificial intelligence in education (AIED) are increasingly subject to scrutiny concerning the fairness of their automated decision-making processes (e.g., auto-grading, problem sequencing, personalization, interventions). However, little attention has been paid to the fairness of the procedures by which these algorithmic decisions are made, i.e., their procedural fairness. This procedural fairness is of critical importance; even seemingly fair decisions can be objectionable to students, teachers, and other stakeholders if the decision-making process itself is unfair. Without a structured framework, AIED applications risk perpetuating biases that disadvantage certain groups, deepen educational inequalities, and marginalize minority voices. Our full-day `Fair4AIED` workshop will take the first steps toward developing a blueprint to systematically integrate algorithmic fairness into AIED research, practice, and adoption. Participants are engaged in short-lightning talks, Q&A sessions, and discussions to explore opportunities to improve fairness and raise awareness of its importance for AIED. Subsequently, small-group discussions followed by a world-café-based session foster a structured, hands-on approach to envision the blueprint's content.

Keywords: Bias · Fairness · Procedural Fairness · Fair AI · Artificial Intelligence in Education · AI Ethics in Education

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 308–314, 2025.
https://doi.org/10.1007/978-3-031-99267-4_40

1 Scope and Relevance

The growing adoption of AI in education is reshaping teaching and learning [7,8,22]. AI is now used in adaptive learning platforms [10], intelligent tutoring systems [18], automated grading [19], and personalization [3], as examples, enhancing education via tailored experiences, efficiency, and process improvement.

Despite these advancements, ensuring fair outcomes in AI for education remains a critical challenge from both research [2,4,6,12] and regulatory[1] perspectives, due to the growing reliance on data-driven decision-making and the vast diversity of learners [9,14,16,21]. For instance, AI systems can amplify biases from training data, leading to unfair outcomes, especially for marginalized groups [1,5,23]. Similarly, AI-powered essay grading may inherit subjective biases from human evaluators [20]. Moreover, the rapid emergence of generative AI applications in education, including decision-making systems that dynamically suggest content and produce feedback [15,17] introduces additional fairness, ethics, and responsibility challenges that remain largely underexplored in current research.

Many studies prioritize overall accuracy and learning outcomes but neglect their impact on varied groups, leaving fairness an open question [22]. When fairness is considered, it is often based on arbitrary decisions or generic technical definitions that lack clear contextualization to education [1,11]. Broadly, questions of fairness often focus on decision makers (e.g., teachers, course designers, researchers), rather than decision recipients (e.g., students) [13].

To bridge these gaps, a community-driven effort is needed to systematize fairness integration into AIED research, practice, and adoption, extending also to emerging domains such as generative AI and AI-driven decision-making systems. The `Fair4AIED` workshop fosters critical discussions on building an operational blueprint to achieve this goal. It engages participants in addressing key themes, including defining fairness, identifying at-risk groups, translating fairness into measurable metrics, mitigating bias, reporting fairness analysis into research, and auditing AIED systems for fairness. By bringing together the community, the workshop aims to set the foundation for establishing best practices to ensure that fairness is consistently and effectively considered in AIED.

The `Fair4AIED` workshop is highly relevant to AIED 2025, as the related community plays a key role in developing intelligent educational tools. Ensuring this community's perspectives on fairness are considered is crucial for a shared direction. Fully aligned with the conference's focus, the workshop fosters collaborative dialogue, contributing to a shared operational framework and strengthening a broader community of practice to systematize fairness in AIED.

2 Organization and Outcomes

The `Fair4AIED` workshop brings together researchers, practitioners, policymakers, and industry professionals whose expertise is essential for embedding fair-

[1] As an example, EU Artificial Intelligence Act (https://artificialintelligenceact.eu/.

ness in AIED systems. In particular, researchers play a key role in designing such systems, making them central to developing fairness-aware methodologies. Ethics and policy specialists shape the regulatory and governance frameworks that define how fairness is implemented, ensuring alignment with societal and legal standards. Industry professionals and edtech developers are responsible for deploying these systems in the real world, where fairness affects individuals.

Table 1. Summary of Workshop Organization.

Session	Description
Morning Part	
Keynote Talk	A presentation highlighting advances and emerging challenges concerning fairness in AIED
Lightning Talks	Short slide-based presentations delivered by participants, based on their submitted abstracts
Abstracts' Discussion	Attendees engage in direct dialogue with presenters to discuss the abstracts' content
Morning Summary	A summary sheet highlighting the key points from the keynote and lightning talks
Afternoon Part	
Opening Briefing	Introduction to the afternoon objectives, structure, and activities, including the use of Slido
Grounding Round(s)	Participants are divided into small groups to analyse thought-provoking prompts
Grounding Report(s)	Each group presents key insights to the broader audience, highlighting multiple viewpoints
World Café Interaction	A structured, rotating discussion format where each table focuses on a section of the blueprint
Closing	Table hosts summarize final insights, followed by a closing session led by the chairs

The workshop is structured into sessions, distributed across the morning and the afternoon (Table 1). The morning part consists of two sessions. In the first session, we invite a keynote speaker to share the current research challenges related to the specific themes of the workshop (e.g., around students' and instructors' perceptions of procedural fairness; metrics, evaluations, and definitions of

procedural fairness in education; effect of procedural fairness in education; key challenges and limitations such as abusing procedural fairness, control of stakeholder voice, and belongings) and potential future directions, followed by Q&A. In the second session, we will have short lightning talks from participants who had the opportunity to submit an abstract in advance of the workshop. Each presenter will have approximately 5 min to present, followed by a dedicated Q&A session. After the talks, each presenter will be seated at a table and other participants will be able to walk around and discuss with presenters about their specific abstracts. At the end of morning part, we offer a brief summary of key issues and works from the keynote and the abstracts presented, which will also serve as a take-home flier to encourage workshop participants to reach out to the organizers and among each other as well as engage in the afternoon part.

In the afternoon part, the workshop includes small-group discussion sessions followed by a world-café-based session to foster a structured, hands-on approach towards the blueprint's drafting. To this end, the organizers first introduce themes and content, significance and relevance, objectives and outcomes, the format and activities planned for the afternoon. `Slido` is used to gather information about the participants and form diverse groups of 4 to 10 participants (resulting in 4–6 groups depending on attendance to the workshop). This introduction ends with thought-provoking prompts that encourage exploration (e.g., "What should an operational framework for fairness in AI for education include, from an interdisciplinary perspective?"). The workshop then covers two sessions. In the first session, participants are divided into small working groups, each tasked with analysing one of the aforementioned prompts from technical, ethical, and practical perspectives. Each group leverages diverse expertise and documents key takeaways—including challenges, solutions, and open questions—after a focused discussion led by a self-assigned moderator. Each group finally shares its key insights with the broader audience, summarizing main points and highlighting diverse perspectives, e.g., using a shared document/slide deck/whiteboard. Additional rounds may be conducted, time permitting, with each group focusing on a different prompt in each round. Then, in the second session, the workshop organizers introduce the World Café methodology, outlining its rules, goals, and structure, aiming to synthesize the insights from the rest of the workshop into an actionable blueprint. Each table, moderated by a host, focuses on a key blueprint section, ensuring guided discussions and continuity as participants rotate. Participants rotate between tables during different rounds: each round includes a recap by the table host, an open discussion, and a last part that aims to draft documentation capturing key takeaways and actionable steps. The workshop will end with each table host summarizing insights, followed by a final part led by the workshop chairs to refine the blueprint and explore post-workshop initiatives.

3 Organizers' Biography

Clara Belitz is a Doctoral Candidate at the University of Illinois Urbana-Champaign. Her research focuses on algorithmic fairness for intelligent tutoring systems, with an emphasis on student voices and experiences. She also researches

surveillance technologies and the appropriate uses of data for personalization. She is also a community organizer.

Nigel Bosch is an Assistant Professor at the University of Illinois Urbana–Champaign, with a joint appointment in Information Sciences and Educational Psychology. His research concerns machine learning, algorithmic fairness, human-centered computing, notably in educational contexts.

Lan Jiang is a Doctoral Student at the University of Illinois Urbana–Champaign. Her current research focuses on developing automated text assessment systems across various contexts and on identifying and reducing bias within these systems. Her broader research interests include understanding student learning behaviors and enhancing learning experiences through natural language processing and machine learning.

Ioana Jivet is a Junior Research Professor leading the Learning Analytics group at CATALPA, FernUniversität in Hagen. Her main research interests focus on providing feedback in higher education via student-facing learning analytics, including designing learning analytics dashboards and generating textual feedback. She is also researching cultural factors that influence the personalization and uptake of learning analytics among students and teachers. She has a MSc in Computer Science from TU Delft and a PhD in Learning Analytics from the Open University of the Netherlands.

HaeJin Lee is a Doctoral Student at University of Illinois Urbana–Champaign. Her research lies at the intersection of human-computer interaction, learning analytics, and explainable AI in educational contexts. Specifically, she focuses on developing and evaluating learning interventions using explainable AI methods in computer-based learning environments.

Badmavasan Kirouchenassamy is a Doctoral Student at Sorbonne University, specializing in the development of automated adaptive feedback systems for online programming education. His research focuses on leveraging reinforcement learning techniques to enhance real-time decision-making. His interests extend also to related areas, including automated error detection in student-written code and feedback generation using information retrieval techniques. His work aims to improve the efficiency and personalization of learning experiences in programming education.

Barbara Leporini is an Associate Professor of Computer Science at the University of Pisa, previously a senior researcher at ISTI-CNR. Her research focuses on Human-Computer Interaction, particularly methods and tools for accessible and inclusive digital interfaces. She has led and participated in numerous national and international projects on accessibility, assistive technologies, and education. She is actively engaged in educational and accessibility policy, contributing to defining accessibility requirements in Italian law. She currently directs a training agency and serves on the Board of Directors of the National Institute for Research, Training, and Rehabilitation (I.Ri.Fo.R.).

Francesca Maridina Malloci is an Assistant Professor at the University of Cagliari, specializing in predictive analytics and recommender systems, with a strong focus on decision-making algorithms in multi-stakeholder contexts, includ-

ing education. She has published extensively in international journals and conference proceedings, contributing to the intersection of AI, fairness, and education. Her current research explores how recommender systems can support fair and transparent decision-making in educational settings. In 2019, she was a visiting scientist at the EURECAT Technology Centre (Spain), collaborating with the Data Science and Big Data Analytics Unit.

Mirko Marras is a Tenure-Track Assistant Professor at the University of Cagliari, specializing in responsible AI, with a strong focus on AI in education and fairness. He has published over 80 papers in top conferences and journals and serves on program committees for major venues, including AIED, EDM, LAK, ICALT, EC-TEL, and UMAP. He has co-chaired multiple workshops, including BIAS (ECIR 2020-2023), FATED (EDM 2022), L2D (WSDM 2021), and RKDE (ECML-PKDD 2023-24). He has been General Co-Chair of ACM UMAP 2024, Blue Sky Track Co-Chair at AIED 2024, Proceedings Co-Chair at EDM 2023-25, and IR for Good Track Co-Chair at ECIR 2024 and 2025. He was a Postdoc at EPFL in 2021.

Daniela Rotelli is a Postdoctoral Researcher at Sorbonne University (France), specialising in learning analytics and educational data mining. She earned her doctoral degree in Computer Science at the University of Pisa (Italy), focussing on data preprocessing and modelling student behaviour in online learning environments. Her research aims to develop data-driven frameworks for understanding temporal learning behaviours, identifying trends, predicting challenges, and supporting fair and effective interventions.

Frank Stinar is a Doctoral Student at the University of Illinois Urbana–Champaign. His research focuses on how normative ethics, algorithmic fairness, and student modeling all interact. Broadly, his research aims to make education more equitable.

References

1. Baker, R.S., Hawn, A.: Algorithmic bias in education. Int. J. Artif. Intell. Educ. 1–41 (2022)
2. Balloccu, G., Boratto, L., Cancedda, C., Fenu, G., Marras, M.: Knowledge is power, understanding is impact: utility and beyond goals, explanation quality, and fairness in path reasoning recommendation. In: European Conference on Information Retrieval, pp. 3–19. Springer (2023)
3. Bernacki, M.L., Greene, M.J., Lobczowski, N.G.: A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose (s)? Educ. Psychol. Rev. **33**(4), 1675–1715 (2021)
4. Boratto, L., Fenu, G., Marras, M., Medda, G.: Practical perspectives of consumer fairness in recommendation. Inf. Process. Manag. **60**(2), 103208 (2023)
5. Burriss, S.K.: An AI ethics primer for educators. Revolutionizing English Education: The Power of AI in the Classroom, p. 179 (2024)
6. Cock, J.M., Bilal, M., Davis, R., Marras, M., Kaser, T.: Protected attributes tell us who, behavior tells us how: a comparison of demographic and behavioral oversampling for fair student success modeling. In: LAK23, pp. 488–498 (2023)

7. Dessí, D., Dragoni, M., Fenu, G., Marras, M., Reforgiato Recupero, D.: Deep learning adaptation with word embeddings for sentiment analysis on online course reviews. Deep learning-based approaches for sentiment analysis, pp. 57–83 (2020)
8. Dessì, D., Fenu, G., Marras, M., Reforgiato Recupero, D.: Leveraging cognitive computing for multi-class classification of e-learning videos. In: The Semantic Web: ESWC 2017 Satellite Events: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28–June 1, 2017, Revised Selected Papers 14, pp. 21–25. Springer (2017)
9. Drachsler, H., et al.: Ethical and privacy issues in the application of learning analytics. In: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, pp. 390–391. ACM, Poughkeepsie New York (2015). https://doi.org/10.1145/2723576.2723642
10. Dutta, S., Ranjan, S., Mishra, S., Sharma, V., Hewage, P., Iwendi, C.: Enhancing educational adaptability: a review and analysis of AI-driven adaptive learning platforms. In: 2024 ICIPTM, pp. 1–5. IEEE (2024)
11. Fenu, G., Galici, R., Marras, M.: Experts' view on challenges and needs for fairness in artificial intelligence for education. In: AIED, pp. 243–255. Springer (2022)
12. Gómez, E., Shui Zhang, C., Boratto, L., Salamó, M., Marras, M.: The winner takes it all: geographic imbalance and provider (un) fairness in educational recommender systems. In: ACM SIGIR, pp. 1808–1812 (2021)
13. Heuer, L.: What's just about the criminal justice system - a psychological perspective. J. Law Policy **13**(1), 209–228 (2005). https://brooklynworks.brooklaw.edu/jlp/vol13/iss1/11
14. Holmes, W., et al.: Ethics of AI in education: towards a community-wide framework. Int. J. Artif. Intell. Educ. **32**(3), 504–526 (2022). https://doi.org/10.1007/s40593-021-00239-1
15. Hu, X., Xu, S., Tong, R., Graesser, A.: Generative AI in education: From foundational insights to the socratic playground for learning. arXiv preprint arXiv:2501.06682 (2025)
16. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
17. Lindsay, E.D., Zhang, M., Johri, A., Bjerva, J.: The responsible development of automated student feedback with generative AI. arXiv:2308.15334 (2023)
18. Mousavinasab, E., et al.: Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. ILE (2021)
19. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artif. Intell. Rev. **55**(3), 2495–2527 (2022)
20. Review, T.P.: The evolution of education: how AI is reshaping grading (2024). https://www.princetonreview.com/ai-education/how-ai-is-reshaping-grading. Accessed 2 Mar 2025
21. Slade, S., Prinsloo, P.: Learning analytics: ethical issues and dilemmas. Am. Behav. Sci. **57**(10), 1510–1529 (2013). https://doi.org/10.1177/0002764213479366
22. Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., Du, Z.: AI in education: A systematic literature review. Expert Systems with Applications (2024)
23. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: evaluating different sources of student data. Int. Educ. Data Mining Soc. (2020)

Multimodal Interactions and Explainable AI for Reflective Physical and Online Learning: MiXai^learn Workshop

Rwitajit Majumdar[1(✉)], Huiyong Li[2], Brendan Flanagan[3], Shin-Ichiro Kubota[1], Soumyadeep Bhattacharya[4], Aditi Kothiyal[5], Prajakt Pande[6,9], Olga C. Santos[7], and Irene-Angelica Chounta[8]

[1] Kumamoto University, Kumamoto, Japan
majumdar@kumamoto-u.ac.jp, kubota@cc.kumamoto-u.ac.jp
[2] Kyushu University, Fukuoka, Japan
li.huiyong.194@m.kyushu-u.ac.jp
[3] Kyoto University, Kyoto, Japan
flanagan.brendanjohn.4n@kyoto-u.ac.jp
[4] Microsoft, Mountain View, USA
soumyadeepb@microsoft.com
[5] IIT Gandhinagar, Gandhinagar, India
aditi.kothiyal@iitgn.ac.in
[6] Southern Methodist University, Dallas, USA
[7] UNED, Madrid, Spain
ocsantos@dia.uned.es
[8] University of Duisburg-Essen, Essen, Germany
irene-angelica.chounta@uni-due.de
[9] Aarhus University, Aarhus, Denmark
prajakt@au.dk

Abstract. This half-day workshop discusses application of explainable AI and multimodal interactions with personal devices, tangible electronics and social robots for fostering reflective learning practices in physical and online learning scenarios. Reflective practices in education have been shown to improve student achievement and prepare students for a sustainable society. Still, learning activities that are across physical and online learning spaces make it difficult to assess reflection skills and support them comprehensively. Recent advances in robotics, sensors, machine learning and generative AI provide opportunities to create technologies to capture and analyze data in a hybrid format across physical and online spaces. The workshop aims to bring together experts from the fields of AI, learning analytics, sensing technologies, and learning sciences to discuss technologies for detecting, understanding, and predicting learning distributed in physical and online spaces, as well as to discuss what is pedagogically desirable, technically feasible, and ethically sound. As a technology context we will demonstrate the functionalities of the LA-ReflecT platform, which is being developed and used as part of this international collaboration. This workshop is important to the AIED community, particularly those developing technologies to capture learning in hybrid spaces or, more generally, complex multimodal student interactions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 315–320, 2025.
https://doi.org/10.1007/978-3-031-99267-4_41

Keywords: Explainable AI · Multimodal interactions · Robotics · Multimodal Learning Analytics

1 Background and Objective

While current online learning management systems often focus on the cognitive skills, there is a lack of data-driven approaches for tracking learning process and provide feedback on learning tasks that are distributed in physical and virtual spaces. This workshop brings together members of a joint international collaboration who are part of different collaborative project funded by JSPS in Japan and European Union to further discuss and explore an innovative integration of AI in a technical infrastructure to enhance learner's critical reflection in the psychomotor and cognitive aspects of learning. This international collaboration proposes the **miXai^learn workshop** to network further with researchers and practioners participating in AIED 2025 to bridge the cyber-physical divide by integrating technologies such as socially assistive robots in the physical space and AI-driven agents in the cyber space and model feedback for reflection both by teachers and learners. The interdisciplinary discussion would focus on the core scientific research question: *How multimodal interaction space with data-informed technologies can support embodied and reflective learning activities?*

2 Relevance and Importance to the AIED Community

This workshop is important to the AIED community because it addresses the challenge of capturing and assessing complex, multimodal student interactions and educational activities distributed in physical and online space. There is limited support for assessing and supporting them, leaving the whole responsibility to the teacher. This workshop aims to foster a discussion that can be beneficial to the AIED community, particularly those working on developing technologies to capture seamless learning activities or more generally, complex and unstructured student interactions in physical and online spaces.

3 Content and Themes

The call for participation welcomes discussions regarding individual researchers' ongoing research aligned to the following topics:

1. Data Collection: Papers exploring the types of data that can be collected during both physical and online reflective learning sessions and the methods for collecting this data.
2. Modeling behaviors during reflective learning sessions
3. Inferences: Studies detecting reflection skills from multimodal log data during physical and online reflective learning sessions.
4. Assessment approaches and ontologies: Research on how to assess students' reflection skills with data, as well as discussions on setting criteria for the assessment of value and fairness in evaluation.

5. Intelligent systems: Papers examining the development and use of intelligent systems to support reflective learning and the challenges faced in implementing these systems.
6. Cultural considerations and other characteristics for personalization: Studies on the role of culture or individual characteristics in the development, testing, and implementation of technologies to support reflective learning.
7. Translation to practice: Research on the barriers and challenges faced in translating research to practice and the strategies for overcoming these obstacles.
8. Ethics, Regulation, and Privacy: Analyses of the ethical considerations for collecting student data during physical and online learning activities and the steps that need to be taken in order to safeguard student privacy. Discussions on the steps that need to be taken to ensure that these technologies are beneficial to students as well as mechanisms to mitigate or exploit the observer effect.

4 LA-ReflecT Platform

LA-ReflecT is a multimodal learning analytics enabled learning platform (Majumdar et al. 2023, 2024c, 2025) that can be linked as external tool to Learning Management Systems (LMS) through LTI. Various case studies that utilized the platform and its related analytics approaches will be shared with the participants. They include tracing online interactions (Majumdar et al. 2024a, 2024b), utilizing gaze traces for group formation (Majumdar et al. 2025), deriving indicators of interactions for characterizing thinking processes (Mishra et al. 2021), modeling movement training and its reflection (Portaz et al. 2024, Santos et al. 2023). The workshop participants will have hands on group activities to explore the platform affordances (Fig. 1) in relation to the workshop theme mentioned above playing the role of teachers as well as students as user of the platform.

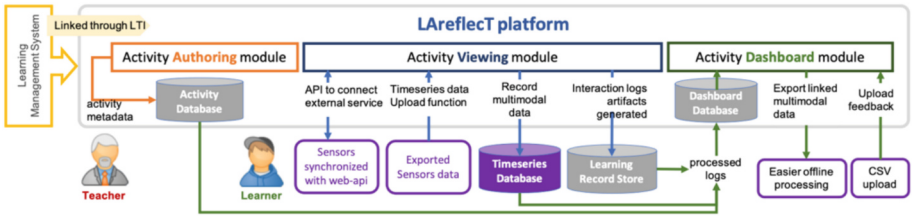


Fig. 1. LA-ReflecT a multimodal analytics enabled learning platform. (Majumdar et al. 2024c).

5 Workshop Logistics

5.1 Intended Audience

We expect participants to be multidisciplinary and work on projects from various perspectives such as learning analytics, instructional design, sensing technologies, and AI. We also welcome participants that conduct research related to project work such as supporting collaboration or inquiry-based education as well as teachers and other practitioners.

We plan to have 10 to 15 participants be part of the workshop, excluding the workshop organizers. Participation in the workshop will be open, which means that it is not mandatory for attendees to submit a paper and make a presentation. The workshop organizers will select presenters based theme that they have mentioned in the participation form, with the goal of ensuring a well-rounded selection of topics.

The organizing team is multidisciplinary and includes experts from the fields of learning analytics, learning science, educational technologies, and sensing technologies.

Call for participation link: https://sites.google.com/view/mixailearn/aied2025.

6 Proposed Pre-Workshop Timeline

Submission of Participation form: May 16
Participants should complete and submit the workshop participation form by May 16, 2025. They can choose to be presenters or participants in the workshop.

Notifications of Acceptance of Participation: June 16
Notification of acceptance will be sent out by June 16, 2025. Accepted presenters will be invited to give an oral presentation at the workshop. We will ask presenters to agree to the recording, archiving, and dissemination of their presentations and question-and-answer sessions.

Sharing Themes for Case-Study Discussion: July 11
The program committee will peer-review and curate the participation form content. This will ensure a well-balanced and focused workshop. We will collate the submitted information in an open-access workshop website as a first step and, along with the interested participants to co-author a journal article, synthesizing the different perspectives that emerged during the workshop.

The workshop will be held on July 26, 2025, from 2 pm - 6pm CET.

7 Format and Activities

We propose a half-day workshop with expert panel, presentations and group activities, in the afternoon. Online participants will also be supported if required. We have provided a detailed tentative schedule in Table 1.

Expert Panel: The expert panel with the members of the organizing committee will initiate the research agenda and the themes that emerged from the submissions and align to their own research focus. Each panelist will briefly highlight the key challenges in that area and possible thought pointer from the perspective of AIED.

Presentations: The format of the presentations will be in the style of lightning talks, with presenters allotted 5 min to present their papers. The presentations will be organized into groups based on similar themes, and after each presentation, there will be a 30-min period for questions and discussions. The organizers will provide guiding questions to facilitate the discussions.

Group Activity: During the workshop, the participants will be organized into smaller groups of 4 to 5 individuals. Each group will be provided with a specific topic related

to the theme of hybrid learning activities, such as the detection of interactions during hybrid learning activities, the evaluation of hybrid learning activities, the automatic support of hybrid learning activities, or the personalization of hybrid learning activities. The groups will then be asked to collaboratively construct a mind map centered around their designated topic. Mind maps are visual diagrams that allow for the representation of concepts and ideas in a non-linear, creative manner. Online participants will form a group via zoom and create the mindmap on Padlet whose link will be shared on the workshop website. Following the construction of the mind maps, we will dedicate time to the sharing of the groups' work and the discussion of their mind maps. The group activity will allow participants to engage in a multidisciplinary discussion of the various challenges and opportunities related to bridging psychomotor and cognitive aspects of learning within the context of AIED.

Table 1. Schedule of activities

Time	Activity
14:00–14:15	Introduction to the workshop
14:15 - 14:45	Expert Panel
14:45 - 15:30	Presentations, Session 1 (3–4 lightning talks)
15:30 - 16:00	Coffee break
16:00 - 16:45	Interaction with LA-ReflecT platform
16:40 - 17:30	Group activities – Review literature and case studies
17:30 - 17:45	Discussion
17:45 - 18:00	Wrap up

Drawing on the diverse expertise of the organizing committee and the participants, this workshop will follow a similar approach as a thematic working group (e.g. Ifenthaler, Majumdar, et al. 2024) to synthesize a journal contribution.

8 Workshop Organizers

The co-organizers are collaborators on the miXai project and different sub groups have conducted various joint workshops in different international conferences. Those workshops and tutorials spanned different topics from Embodied Cognition workshops in ICCE (ICCE 2021–2024), Learning Analytics Data Challenges at LAK (LAK 2018–2025), UMAP tutorial on Modeling Intelligent Psychomotor Systems (Portaz, Garcia, Majumdar and Santos, 2024), LA-ReflecT workshop (Majumdar et al. 2023b, 2024d). For the AIED conference in 2023, organized a tutorial on Designing, Building and Evaluating Intelligent Psychomotor systems (Santos et al., 2023) and a Workshop on Automated Assessment and Guidance of Project Work (Abou-Khalil et al., 2023).

Acknowledgments. This workshop is partially supported by JSPS Grant-in-Aid for Joint International Research B (24KK0051). This collaboration is part of several research projects: HumanAID

(TED2021-129485B-C41) by MCIN/AEI and NextGeneration EU/PRTR, JSPS KAKENHI JP24KK0051, and Moodtion (UNED-IUED).

References

1. Abou-Khalil, V., Vargo, A., Majumdar, R., Magno, M., Kapur, M.: Automated assessment and guidance of project work. In: AIED 2023, vol. 1831. Springer, Cham (2023).https://doi.org/10.1007/978-3-031-36336-8_10
2. Ifenthaler, D., et al.: Artificial intelligence in education: implications for policymakers, researchers, and practitioners. Technology, Knowledge and Learning (2024). https://doi.org/10.1007/s10758-024-09747-0
3. Majumdar, R., Gatare, K., Warriem, J. K.: LA-ReflecT: Multimodal Learning Analytics of Micro-learning. T4E 2023 workshop (2023b).
4. Majumdar, R., Liang, C., Ocheja, P., Li, H.: A cooperative learning framework with joint attention and interaction data in the LA-ReflecT platform. In: ACM Symposium on Eye Tracking Research & Applications (ETRA) (2025). https://doi.org/10.1145/3715669.3726825
5. Majumdar, R., Ocheja, P., Li, H., Warriem, J.M.: LA-ReflecT Dashboard: Scaffolding Reflection on Learning with Teacher Prompts & LLM Interactions. T4E 2024 Tools track (2024d)
6. Majumdar, R., Prasad, P., Sane, A.: Mining epistemic actions of programming problem solving with Chat-GPT. In: Educational Data Mining Conference 2024 (2024a)
7. Majumdar, R., Prasad, P., Kadam, K., Gatare, K., Warriem, J.M.: LA-ReflecT: a platform facilitating micro-learning and its multimodal learning analytics. In European Conference on Technology Enhanced Learning (pp. 731–735). Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-42682-7_69
8. Majumdar, R., Singh, D., Rajendran, R., Narayanan, S., Gatare, K.: LA-ReflecT platform's affordances for distributed multimodal learning analytics. In: CollabTech 2024 (2024c)
9. Majumdar, R., Singh, D., Chen, M.R.A.: Designing structured reflections for guiding learners' interactions with generative AI. In: CELDA 2024 (2024b)
10. Mishra, S., Majumdar, R., Kothiyal, A., Pande, P., Warriem, J.M.: Tracing embodied narratives of critical thinking. In: International Conference on Artificial Intelligence in Education, pp. 267–272. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-78270-2_4
11. Olga C. Santos. 2016. Training the Body: the potential of AIED to support personalized motor skills learning. Int. J. Artif. Intell. Educ. **26**(2), 730–755 (2016). https://doi.org/10.1007/s40593-016-0103-2
12. Portaz, M., Garcia, P., Majumdar, R., Santos, O.C.: Mastering Mind and Movement. ACM UMAP 2024 tutorial on modeling intelligent psychomotor systems (M3@ ACM UMAP 2024). In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, pp. 9–12 (2024)
13. Santos, O.C., Portaz, M., Casas-Ortiz, A., Echeverria, J., Perez-Villegas, L.F.: Designing, Building and Evaluating Intelligent Psychomotor AIED Systems (IPAIEDS@AIED2023). In: Wang, N., Rebolledo-Mendez, G., Dimitrova, V., Matsuda, N., Santos, O.C. (eds) in AIED 2023. Communications in Computer and Information Science, vol. 1831. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36336-8_14

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 Workshops, CCIS 2592, pp. 321–327, 2025.
https://doi.org/10.1007/978-3-031-99267-4

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