Innovation and Responsibility in AI-Supported Education Workshop at the 39th Annual AAAI Conference

Debshila Basu Mallick DEBSHILA@RICE.EDU
SafeInsights, Rice University

Jill Burstein JILL@DUOLINGO.COM
Duolingo

Simon Woodhead SIMON.WOODHEAD@EEDI.CO.UK
Eedi, United Kingdom

James Sharpnack JSHARPNA@GMAIL.COM
Duolingo

Muktha Ananda MUKTHANANDA@GOOGLE.COM
Google

Zichao Wang JACKWA@ADOBE.COM
Adobe Research

The Innovation and Responsibility in AI-Supported Education (iRAISE) workshop at the 39th Annual AAAI Conference served as a vital convening for the AI for Education research community, convening at a time when the field is experiencing unprecedented growth, largely fueled by advancements in large language models (LLMs) and other Generative AI (GenAI). The workshop gathered a diverse set of professional backgrounds, including researchers, graduate students, postdocs, industry representatives, and a few educators. Amidst the often-polarized discourse surrounding AI's role in education, the workshop's clear objective was to foster a principled, balanced exploration aimed at answering the fundamental research question: ***In what tangible ways can AI make a positive impact on education?***

Our half-day workshop included a keynote, three invited speakers, and an expert panel. Of the 35 submissions, we accepted 21 full and short papers to include in the workshop proceedings. The 10 highest-scoring submissions were included for 90-second Poster Spotlight presentations.

The **Poster Spotlight** session highlighted a remarkable breadth of innovative research in AI for education. We saw novel frameworks aimed at enhancing AI-generated narratives for children by improving consistency (Shahriyar et al.), a known flaw in current systems, and yet success in generating supplementary learning materials (Liou et al.). We saw a multimodal GenAI learning analytics framework designed to capture biometric and behavioral data for deeper student insights (Becerra et al.) as well as multimodal student support (Sun and Tai). Researchers shared a scalable AI tutoring framework for low-resourced communities (Hevia et al.). The researchers presented AI platforms designed to enhance teacher expertise in instructional design, focusing on teacher control and partnership, and

not to replace them (Shahriyar et al.; Sucholutsky et al.; Heady and Thais). One project focused on simulating students to assess their knowledge with adaptive interview assessment, demonstrating benefits of the adaptive over a fixed policy for question selection (Ion et al.). Significant attention was paid to assessment and feedback for personalized learning (Silva and Costa; Kakarla et al.; Ozturk et al.). Multiple projects explored the use of LLMs for generating automated scores and feedback on programming assignments and student essays /citephou25,latif25,buckley25, often incorporating linguistic features to improve accuracy. These studies revealed that although LLMs were effective in handling aspects such as syntactic errors, they continue to demonstrate weaknesses, such as making logical errors. This highlights a recurring theme: while the potential of LLMs to have a positive impact on assessment and instruction is clear, their current limitations require continuous evaluation.

The workshop explored the current state and future of Generative AI (GenAI) for learning and assessment, emphasizing responsible innovation, with invited speakers representing various areas of expertise. Lisa Wang (Google DeepMind) presented the evaluation-driven LearnLM project, designed to enhance Gemini through pedagogical grounding and real-world deployment. The project uses a hierarchical evaluation framework that includes automated metrics, human evaluations, and efficacy studies. Keynote speaker Venu Govindaraju (University at Buffalo) emphasized that context is a key human advantage over AI, noting that AI systems often struggle in its absence. He highlighted the importance of accurately capturing data rather than interpreting it, especially in sensitive areas such as dysgraphia analysis. Lydia Liu (Princeton University) advocated shifting the focus from predictive accuracy to how AI systems can enhance human expertise. She proposed using causal models to better understand this interaction. Diane Litman (Univ. Pittsburgh) presented eRevise+RF as a case study in responsible innovation, detailing how human-centered design, privacy, and transparency guided the development of an automated writing feedback tool. Together, these presentations mapped pathways for developing and evaluating effective, ethical tools supporting AI for education that prioritize human context, expertise, and responsible practices. Finally, our expert panel with legal experts, including Amelia Vance, Ravit Dotan, Jeff Knight, and moderated by Jeremy Roschelle, emphasized the ethics of GenAI use, discussing trust, privacy, and responsible deployment, offering practical advice on stakeholder engagement and policy navigation during the panel discussion on **"Cultivating Trust for Generative AI Applications Among Educational Communities"**.

From a research standpoint, the workshop emphasized critical challenges that demand rigorous investigation:

- **Human-AI Collaboration.** A dominant theme emerging from the presented research is the critical importance of collaboration between humans and AI (Li et al.; Sun and Tai). Rather than aiming for full automation, the most promising and responsible path involves designing AI systems explicitly intended to augment and collaborate with human educators and learners. Researchers are actively working to identify and value forms of human expertise, such as nuanced advising strategies or contextual understanding crucial for supporting students with specific needs that are difficult, if not impossible, to fully encode algorithmically. Successful real-world implementations, that Lydia Liu shared, like Georgia State's predictive analytics system,

underscore the value of a "human-first" approach, coupling technology with significant investment in human capital (e.g., hiring more advisors).

- **Efficacy Beyond Efficiency.** The gap between efficient task completion and effective learning gains remains a primary concern. Conversely, preliminary evidence suggests that AI tools designed to support human tutors may yield more positive student outcomes (Sucholutsky et al.; Ye and Wu). This necessitates robust efficacy studies, such as randomized controlled trials or causal analyses as proposed by Lydia Liu, to move beyond correlational findings.

- **Model Reliability and Validity.** LLMs, despite their power, exhibit limitations. Research presented detailed struggles with logical reasoning in programming feedback (where 37% of generated messages contained errors)(Silva and Costa), hallucination, inconsistency, multilingual settings (Berman et al.), and difficulties processing nuanced human inputs like children's handwriting or accommodating conditions like dyslexia, where human contextual understanding remains superior. The risk of incorrect AI guidance misleading students is nontrivial.

- **The Evaluation Quagmire.** Standard benchmarks (e.g., GSM-8K for math) are becoming saturated and may not accurately reflect true model capabilities or rank models reliably, especially when examining performance on challenging, discriminative questions (Castleman et al.; Andres and Whitmer). This highlights the need for more sophisticated evaluation frameworks, incorporating methods like Item Response Theory (IRT) and computational psychometrics, developing scenario-based human evaluations, and always grounding automated metrics in careful human annotation.

- **Fairness, Bias, and Transparency.** Empirical studies, such as those on remote proctoring, reveal existing human biases (e.g., nationality-based in-group leniency) (Belzak et al.). While AI detection tools showed potential in promoting fairer decisions in this context, the inherent biases within AI models themselves require continuous scrutiny. Furthermore, the "black box" nature of many AI systems remains problematic, driving research towards greater transparency, explainability, and user control, including customization to specific pedagogical needs (Latif et al.; Li et al.; Belzak et al.). Legacy models like BERT sometimes even outperform contemporary LLMs on nuanced tasks like equity assessment, cautioning against assuming newer is always better (Kakarla et al.).

- **AI for Education is increasingly multimodal.** This integration of diverse data types—text, images, audio, video, and even biometric inputs—opens exciting avenues for novel research. Investigators are already exploring how combined modalities can foster more holistic, personalized, and engaging learning experiences (Sun and Tai), develop sophisticated AI tutors capable of nuanced, multisensory interaction, and design new forms of assessment that capture a richer spectrum of student competencies. This direction also calls for research into the unique ethical considerations and practical challenges of effectively and responsibly harnessing multimodal generative AI in educational settings.

1. Role of researchers in the path forward

The AI for education researchers working alongside practitioners and AI and edtech developers are instrumental for the responsible and effective use of AI for instructional purposes. Researchers must leverage robust experimental designs, including control conditions, to establish causal links between AI interventions and learning outcomes. Using evidence-based approaches to actively investigate and transparently report the limitations of AI systems alongside their capabilities is critical for developing pedagogically aligned AI to improve educational outcomes. Starting with a human-centered design approach, engaging all stakeholders (students, teachers, administrators, policymakers) in co-design processes from the outset, grounded in authentic use cases, will ensure that tools and technologies and usable and useful in various learning contexts. Finally, to address issues of fairness, privacy, and transparency, we must bring existing policy frameworks (like Family Educational Rights and Privacy Act (FERPA) and Children's Online Privacy Protection Act (COPPA)) to the challenges and needs of this day and age.

2. Conclusion

The iRAISE workshop demonstrated that the path forward requires a concerted effort. It involves rigorous research into model capabilities and limitations, the development of better evaluation methodologies, and a commitment to human-centered design that prioritizes pedagogical soundness and ethical considerations. The emphasis on grounding AI development in real-world deployments and user feedback, as exemplified by Google's work on LearnLM and Gemini, provides a valuable model. Collaboration between AI researchers, educators, learning scientists, policymakers, and industry is not just beneficial, but essential. As we continue to explore the potential of AI to reshape education, the insights shared at iRAISE serve as a crucial reminder: true progress lies not just in technological advancement but in our collective ability to wield these powerful tools wisely, ethically, and always in service of meaningful learning.

Steering committee. Muktha Ananda (Google), Debshila Basu Mallick (OpenStax-Rice University), Jill Burstein (Duolingo), James Sharpnack (Duolingo), Jack Wang (Adobe), Simon Woodhead (Eedi, UK).

Acknowledgements. We thank the conference organizers, all our speakers and presenters, and attendees for participating in the workshop and making it a success. We sincerely thank Eedi for sponsoring the travel scholarships that enabled two student participants to attend the workshop and present their work.

References

J. M. Alexandra L. Andres and John Whitmer. The levi training hub: Evidence-based evaluation for ai in education. pages 202–211.

Alvaro Becerra, Roberto Daza, Ruth Cobos, Aythami Morales, and Julian Fierrez. M2lads demo: A system for generating multimodal learning analytics dashboards. pages 141–145.

William Belzak, Jill Burstein, and Alina A. von Davier. Evaluating fairness in ai-assisted remote proctoring. pages 125–132.

Shmuel Berman, Yuval Kansal, and Lydia Liu. Facts do care about your language: Assessing answer quality of multilingual llms. pages 238–244.

Jane Castleman, Nimra Nadeem, Tanvi Namjoshi, and Lydia Liu. Rethinking math benchmarks: Implications for ai in education. pages 66–82.

Ashley Heady and Savannah Thais. Ai awareness survey of educators. pages 245–249.

Juan Segundo Hevia, Facundo Arredondo, and Vishesh Kumar. Towards an efficient, customizable, and accessible ai tutor. pages 250–254.

Michael Ion, Sumit Ashana, Fengquan Jiao, Tianyi Wang, and Kevyn Collins-Thompson. Adaptive knowledge assessment in simulated coding interviews. pages 260–262.

Sanjit Kakarla, Conrad Borchers, Danielle R. Thomas, Shambhavi Bhushan, and Kenneth R. Koedinger. Comparing few-shot prompting of gpt-4 llms with bert classifiers for open-response assessment in tutor equity training. pages 133–140.

Ehsan Latif, Yifan Zhou, Luyan Fang, and Xiaoming Zhai. Efficient multi-task inference with a shared backbone and lightweight task-specific adapters for automatic scoring. pages 212–220.

Hongming Li, Yizirui Fang, Shan Zhang, Seiyong M. Lee, Yiming Wang, Mark Trexler, and Anthony F. Botelho. Arched: A human-centered framework for transparent, responsible, and collaborative ai assisted instructional design. pages 94–104.

Yow-Fu Liou, Yu-Chien Tang, and An-Zi Yen. Stay hungry, stay foolish: On the extended reading articles generation with llms. pages 230–237.

Aylin Ozturk, Robin Schmucker, and Tom Mitchell. Enhancing learning outcomes within a large-scale online learning system through ai-powered feedback. pages 255–259.

Akib Shahriyar, Radwa Hamed, E. Margaret Perkoff, Mostafa Aboelnaga, and Alya Azab. Bibliosmia: Hyper-personalized consistent stories for enhanced social emotional learning. pages 105–115.

Priscylla Silva and Evandro Costa. Assessing large language models for automated feedback generation in learning programming problem solving. pages 116–124.

Ilia Sucholutsky, Katherine M. Collins, Maya Malaviya, Nori Jacoby, Weiyang Liu, Theodore R. Sumers, Michalis Korakakis, Umang Bhatt, Mark Ho, Joshua B. Tenenbaum, Zachary A. Pardos, Adrian Weller, and Thomas L. Griffiths. Representational alignment supports effective teaching. pages 146–173.

Edward Sun and LeAnn Tai. Multitutor: Collaborative llm agents for multimodal student support. pages 174–190.

Patrick Peixuan Ye and Shirley Wu. Can llms teach human learners to understand concepts through analogies? pages 191–201.