Using Learning Analytics to Unpack Human–AI Roles in Generative AI Math Tutoring*

Lynne Harden[1*], Akila Raoul[1], Rebecca McGill-Wilkinson[1], Amber Wang[1], Katie Boody Adorno[1], and John Scianimanico[2]

[1] *Leanlab Education, 1475 Walnut St #26, Kansas City, Missouri, 64106, United States of America*
[2] *NewSchools Venture Fund, 2648 International Blvd Ste 301, Oakland, California, 94601, United States of America*

Abstract

AI-powered tools are becoming more prevalent in K-12 education, and questions about the evolving roles of humans and AI in supporting student learning have become increasingly urgent. This work-in-progress paper presents early findings from an ongoing mixed-methods study exploring the potential of generative AI to enhance high-impact mathematics tutoring and provides an opportunity to examine the evolving roles of humans and AI in supporting student math learning and motivation, engagement, and persistence. We evaluate seven AI-powered math tutoring tools–four student-facing tools, and three tutor copilots–using a novel, multi-phased research design that integrates LLM-as-a-judge benchmarking with iterative rapid cycle evaluations and controlled impact studies. Preliminary results indicate that the tools effectively support students and tutors with step-by-step instruction yet currently fall short in providing motivational and metacognitive support needed to foster deeper understanding. Combined with literature demonstrating the importance of student-teacher relationships in student academic achievement and engagement, this suggests that hybrid human-AI tutoring models may be the optimal solution for maximizing student math outcomes. The final study phase, a set of quasi-experimental trials in spring 2026, will further assess tool effectiveness in authentic learning environments and refine the emerging framework for human–AI collaboration in K-12 tutoring.

Keywords

artificial intelligence (AI) in education, human-AI collaboration, iterative codesign, high-impact tutoring, intelligent tutoring systems (ITS), mathematics education

1. Introduction

With the rise of large language models (LLMs), artificial intelligence (AI) is increasingly prevalent in classrooms and other educational settings. Recent reports indicate that as many as 85% of teachers and 86% of students in the United States used AI during the 2024-2025 school year [1]. AI integration in education shows promise in personalized learning and individual instruction [2], student engagement [3] and equity and accessibility [4]. However, some worry that widespread AI use in education poses risks for academic integrity [5], data privacy, and the development of critical thinking skills [6]. Inarguably, AI in education is a complex matter that warrants continued research in a variety of contexts.

1.1. Humans vs. AI in education

While the research into optimal AI integration in education is not conclusive, it is evident that AI has the potential to redefine the role of humans–especially teachers–in maximizing student learning and development and transforming the way teachers and classrooms operate. Should AI serve as mainly a supporting tool for teachers, or as a co-teacher, shifting the role of the human

Joint Proceedings of LAK 2026 Workshops, co-located with the 16th International Conference on Learning Analytics and Knowledge (LAK 2026), Bergen, Norway, April 27 – May 1, 2026

[1*] Corresponding author.
lynneharden@leanlabeducation.org (L. Harden); akilaraoul@leanlabeducation.org (A. Raoul); rebeccamcgillwilkinson@leanlabeducation.org (R. McGill-Wilkinson); amberwang@leanlabeducation.org (A. Wang); katie@leanlabeducation.org (K. Boody Adorno); jScianimanico@newschools.org (J. Scianimanico)
0000-0002-0505-6762 (L. Harden); 0000-0003-0978-7236 (A. Raoul); 0009-0004-5875-6772 (R. McGill-Wilkinson); 0000-0003-1328-2134 (A. Wang)

teacher from content deliverer to motivational coach or context connector? In what circumstances, if any, should AI handle parts of the teaching process—like providing direct instruction or facilitating independent practice? Different perspectives on these various human-AI scenarios generally center around the democratization of education and closing equity gaps, personalized and adaptive learning, workload and time management, cost and scalability of high-quality interventions, algorithmic bias, and socio-emotional learning and development. For example, AI has the potential to help close learning gaps in providing scalable, cost-effective, personalized, and adaptive learning to more students while saving teachers time [7]; however, if the digital divide remains between low-income communities and more resourced schools and districts, AI in education could further widen learning and achievement gaps [8].

1.2. Humans vs. AI in high-impact tutoring and the role of teacher-student relationships

Tutoring is a compelling context for exploring the role of AI and humans in education. Research has consistently shown supplemental high-impact tutoring interventions to be highly effective in improving student learning outcomes in various subjects [9] through a personalized learning approach when conducted by highly trained and competent tutors. The inherent longstanding problem with using high-impact tutoring as a larger solution to educational and equity gaps is in its accessibility, scalability, and cost, and AI has the potential to address these issues. Intelligent tutoring systems (ITS), or computer programs that provide personalized instruction to students, have been shown to be effective learning tools across a variety of subject domains [10], particularly when the ITS tailors the instructional content based on an individual student's prior knowledge, learning style, and needs [11].

Despite evidence for the effectiveness of ITS, questions remain regarding the distributed roles of ITS, AI tutors, and human tutors in supporting learning. Indeed, there is evidence that in some cases a hybrid human-AI tutoring model is more effective than AI tutors alone [12], suggesting an integrated human component may enhance the benefits of AI tutors. In their Framework for High-Impact Tutoring, Robinson and Loeb [13] note that high-impact tutoring programs tend to include a stated focus on cultivating tutor-student relationships, and that many of these programs cite a positive tutor-student relationship as a critical feature. Indeed, there is widespread evidence that high-quality positive student-teacher relationships and interactions in learning environments more generally are beneficial to students. When students feel safe and supported in their learning context and a strong connection with their teachers, they are more motivated to learn and persevere through difficulties [14]. Supportive teacher-student relationships can positively impact both student academic achievement [15, 16] as well as student self-esteem [17], an important indirect component of student performance. Within the math learning context specifically, students have demonstrated greater motivation to learn mathematics when provided with teacher support of self-perception and self-determination [18]. Conversely, students risk experiencing harm through synthetic relationships built with AI [19], as has already been observed in some social media contexts [20].

The positive impacts demonstrated by positive student-teacher relationships and the potential risks associated with synthetic relationships built with AI thus raise compelling questions, such as:

1. If a positive student-teacher relationship is an essential component of student achievement and growth, can AI replicate it in such a way as to elicit the same level of motivation and produce similar academic and durable skill outcomes in students? Or are the benefits that students get from a positive relationship with their teacher only achievable through direct human interactions? Additionally, are the potential risks associated with synthetic relationships built with AI able to be mitigated within an educational context, or are the risks not worth the rewards?

2. Is the ultimate role of AI in education to help students practice and reinforce concepts and skills using the lower levels of Bloom's taxonomy while human teachers or tutors provide the necessary motivation, encouragement, and metacognition necessary for long-term retention and knowledge building?

In this paper, we present a study in progress to examine the potential of AI-powered tutoring for math learning and the role of humans–teachers and tutors–in maximizing student outcomes. This work-in-progress paper contributes (1) a multi-phase research design integrating LLM-as-a-judge with human rapid cycle evaluation, (2) early findings on feasibility, engagement, and scaffolding dynamics of seven AI-powered math tutoring tools, and (3) emerging insights into the complementary roles of humans and AI in high-impact tutoring.

2. Methodology

Leanlab Education, in partnership with NewSchools Venture Fund and The Learning Agency, is exploring the impacts and potential of seven different generative AI-powered math tutoring solutions (Table 1) on K-12 student and teacher outcomes. The team is interested in learning if an iterative codesign research approach utilizing both LLM-as-a-judge and human evaluations leads to the development or refinement of AI-powered math tutoring tools that improve student learning, student math motivation, engagement, and persistence, and teachers' or tutors' experience teaching math. Additionally, the diverse feature sets and use cases of the generative AI-powered math tutoring solutions offer an opportunity to examine the role of humans and AI in high-impact tutoring. These questions will be explored through a three-phase mixed methods research design:

1. In the fall of 2025, The Learning Agency utilized an LLM-as-a-judge approach to assess the AI component of each tool on its mathematical accuracy, pedagogical quality, and equity and fairness by evaluating at least 100 interaction outputs from each tool. LLM-as-a-judge evaluators were provided with evaluation guidelines for each tool that were generated by LLMs using research-backed math intervention practice guides, tool-specific context, and human (teacher) input. In addition, Leanlab Education researchers conducted an audit of each tool, assessing their usability, interoperability, accessibility, and existing evidence base.
2. In the fall of 2025, following the initial LLM-as-a-judge benchmark evaluations and audits, Leanlab Education conducted an eight-week rapid cycle evaluation study with each tool. In each cycle, a unique set of five elementary, middle, and/or high school human teachers or tutors tested each tool and provided feedback (35 teachers or tutors in total). Tool developers used the feedback to make iterations and improvements to their tool, which were then tested again in the subsequent cycle. A total of three feedback cycles were conducted. The LLM-as-a-judge and rapid cycle evaluations provided an opportunity for the tool developers to ready their AI-powered math tutoring tools for the third and final phase of research, a controlled efficacy study.
3. Between January and April 2026, Leanlab researchers will utilize a quasi-experimental methodology to explore the relationship between regular tool use in a classroom or other learning environment and student math learning outcomes, as well as student math motivation, engagement, and persistence. They will also explore the feasibility of implementing each tool in an authentic learning environment. A unique set of five or more treatment group teachers or tutors will implement each tool with students in their classrooms or tutoring environments, while a similar number of control group teachers or tutors will teach math "business-as-usual" without using the same generative AI-powered math tutoring tools. Student pre- and post-assessment data as well as motivation, engagement and persistence survey data from these teachers and tutors will be compared to

data from the similar sample of students who will not be exposed to the same AI-powered math tutoring tools.

Table 1
Seven generative AI-powered math tutoring tools involved in iterative codesign and efficacy research led by Leanlab Education

Tool Name	Description
Tutor Assist and Tutor Feedback by Step Up Tutoring	Tutor Assist and Tutor Feedback is a two-in-one AI-powered tutor copilot tool that provides math tutors working with 2nd-6th graders with real-time tutoring assistance (Tutor Assist) as well as feedback and coaching after tutoring sessions (Tutor Feedback).
ALTER-Math	ALTER-Math uses an AI teachable agent to help students in grades 6-8 learn math through a peer teaching model.
Goblins	Goblins is a multimodal AI math tutor that lets kids in grades 5-12 speak naturally, draw freely, and get instant help with solving math problems.
Snorkl	Snorkl is a multimodal AI math tutor that asks K-12 students to solve math problems and explain their thinking verbally.
Sophia by CitySchools	Sophia is an AI-powered tutor copilot tool that a math tutor for grades 3-8 can use to prepare for tutoring sessions, get help during tutoring sessions, and reflect on completed tutoring sessions.
Talk Tree and JIA by the University of Colorado Boulder	Talk Tree and JIA comprise a two-in-one AI-powered math tutor copilot tool for grades 5-8 that can be used to build structure during a tutoring session (Talk Tree) and provide feedback to a tutor after a tutoring session (JIA).
Thinkverse	Thinkverse is a personalized AI tutor that guides students in grades 3-12 through math challenges, utilizing assessment data to provide customized learning pathways.

The seven AI-powered math tutoring tools that will be examined in this multi-phase study were chosen for their robust foundation in the learning sciences and tutoring literature, potential for improving student outcomes using diverse pedagogical methods and approaches, and innovative application of artificial intelligence to authentic learning environments. Four of the tools are student-facing tools in which students interact directly with an AI tutor, while the remaining three are tutor copilot tools. In these cases, tutors use the AI-powered tools either to prepare for tutoring sessions with students, get help or guidance during tutoring sessions, and/or receive feedback on completed tutoring sessions. While the primary focus of this research collaboration is to assess the impact of codesigned generative AI math tutoring tools on student outcomes, the two distinct tool groupings (student-facing vs. tutor copilot) offer a secondary opportunity to examine and compare the role of humans–specifically, teachers and tutors–and AI in supplementary math learning.

As the tools are at various stages of development–Step Up Tutoring's tools, Thinkverse, ALTER-Math, and Snorkl are all fairly well developed, Goblins is somewhat well developed, and Talk Tree/JIA and Sophia are still in development–the goal of the first two phases of research is to prepare each tool for classroom implementation and efficacy evaluation, while acknowledging the variation in developmental stages of each product.

3. Preliminary results

This section outlines preliminary findings from the LLM-as-a-judge benchmark evaluations for five of the seven AI-powered math tutoring tools (excluding Talk Tree/JIA and Sophia), research audits for all seven tools, and two of the three rapid cycle evaluations.

3.1. Implementation feasibility

While all of the student-facing tools incorporate various multimodal learning components, such as allowing students to explain themselves verbally or use drawings or graphs to express their mathematical thinking, the human evaluators (teachers and tutors) in the rapid cycle evaluations expressed concerns over the feasibility of students utilizing the multimodal components in a traditional classroom environment. For example, 20-30 students all speaking to an AI tutor at the same time in the same classroom (using, for example, Goblins or Snorkl) could lead to problems with the AI's understanding of a single student in a sea of voices. Additionally, teachers and tutors expressed concerns over students' ability to draw mathematical pictures or diagrams on a computer when they are used to drawing pictures or diagrams with pencil and paper.

3.2. Motivation, engagement, and persistence

The four student-facing tools use a variety of strategies for engaging students. For example, the ALTER-Math teachable agent asks follow-up questions to the human learner to keep them engaged and moving forward in the problem-solving process. It also awards badges to students for persisting through and solving problems and displays a leaderboard to students so they can see how they rank compared to their classmates. The Goblins AI tutor provides students with humorous and light-hearted feedback that incorporates generationally relatable slang and pop culture language (e.g., "rizz"). Despite these intentional motivation and engagement strategies, teachers explicitly expressed concerns over three of the four tools' abilities to capture and maintain student engagement and motivation. Teacher evaluators of Snorkl and ALTER-Math noted that the AI-generated feedback from both tools contains too much text, and thus is not conducive to keeping students engaged. Teachers assessing Goblins and ALTER-Math also felt that incorporating more gamified components or incentives (e.g., giving students the ability to earn points from solving problems that can be traded in for virtual prizes) would provide more effective motivation for students to stay engaged. While gamification can hook students initially, ultimately, teacher and tutors recognized that long-term engagement is about implementing accountability measures, clear expectations, systems for minimizing student frustration, and making student progress visible.

3.3. Micro-scaffolding vs. macro-scaffolding

The five tools with LLM-as-a-judge benchmark results were all rated highly on mathematical accuracy and can provide helpful step-by-step or moment-to-moment support to students and tutors (micro-scaffolding). However, the student-facing tools in their current form fall short in macro-scaffolding techniques such as helping students reflect on their own thinking, understand why they solved a math problem the way they did, or think about how they could apply the same problem–solving techniques to other scenarios. Most student-facing tools also do not adequately recap, summarize, or reinforce what students learned at the end of a tutoring session. Finally, the student-facing tools frequently struggle to help students who stall out in their thinking, become frustrated, or do not know where to start. Students in these situations can find themselves in unproductive conversational loops in which the AI tutor is unable to provide the appropriate adaptive scaffolding to help them make forward progress.

4. Discussion

The preliminary findings from the first two phases of this research study offer compelling evidence for why the division of labor between humans and AI in high-impact tutoring is a critical question that should be further examined. Across the LLM-as-a-judge evaluations, the AI tutors and tutor copilots consistently demonstrated strong mathematical accuracy and effective step-by-step procedural guidance. This reflects the early promise of AI integration in education: the ability to support individualized practice at scale and scaffold students through moment-to-moment problem solving. In relation to Bloom's taxonomy, the tools support remembering, understanding, and applying, which are important foundations for math learning and proficiency. However, the same tools demonstrated a more limited capacity for engaging students or guiding tutors in higher order cognition, such as evaluating and explaining a problem-solving process, drawing connections between ideas, or reflecting on one's own learning–skills that help build solid conceptual understandings.

Additionally, the educator focus on student motivation and engagement in the rapid cycle evaluations for the student-facing tools validated their importance in math learning and shed light on potential shortcomings in the tools' abilities to sustain student attention and engagement. While the tools attempt to approximate motivational dynamics that are typically offered by human teachers or tutors, they ultimately fall short, and are not yet perceived by teachers as supporting long-term engagement. This initial finding is compelling in that it is consistent with the robust literature on the importance of student-teacher connections and relationships in student motivation, engagement, and persistence.

Teachers raised practical concerns about implementing multimodal AI tutoring tools with a group of students at the same time, particularly voice-based modalities. This illustrates the potential real-world limitations that could come with shifting tutoring towards fully AI-driven models, at least in a traditional classroom/group learning context. It also points to the need for foundational shifts in how classrooms operate to support emerging technologies like AI tutoring tools. These feasibility challenges in conjunction with the current limitations of AI tutors with motivation and metacognition suggest that an AI-human hybrid model in which AI both augments instruction and provides teachers with indicators to target support may be most effective for optimizing student outcomes.

The results presented in this paper are preliminary findings from a multi-phase study in progress. In the spring of 2026, seven individual quasi-experimental design studies will be conducted to test the feasibility and effectiveness of each of the seven AI-powered tutoring tools, and the question of the optimal role of humans (i.e., teacher or tutors) and AI in high-impact tutoring will be further examined. Initial results from this last phase of research will be available between April and June 2026. At that time, we will also examine the effectiveness of the iterative codesign and hybrid human-AI (LLM-as-a-judge) methodologies employed in this research as a potential replicable blueprint for developing AI-powered educational tools that are pedagogically robust, contextually feasible, and responsive to real-world constraints.

As AI adoption in schools accelerates, understanding the optimal integration of human and AI to support both teachers and students will be essential for improving educational outcomes and closing learning gaps. This work-in-progress contributes to an emerging body of scholarship that seeks to ground AI innovation in evidence and the realities of authentic learning environments.

5. Limitations

The preliminary results presented in this paper are part of a work-in-progress and thus should not be over-interpreted without the additional context that will be provided by the final research phase data. Additionally, the results from the rapid cycle evaluations should be interpreted with caution due to the small sample size of teacher/tutor evaluators.

The LLM-as-a-judge model for evaluating each tool's AI interactions, which has been shown to be an effective evaluation method, was utilized for its cost-effectiveness and speed; however, the method was not subjected to a robust comparison (e.g., interrater reliability) with human evaluators.

Finally, while we aim to explore the important question of the role of humans and AI in effective educational interventions like high-impact tutoring in this multi-phase research study, this study was not specifically designed to answer this research question. Thus, it may not yield conclusive results as to the most effective way to integrate AI into math tutoring interventions; however, it is likely to provide important insights to the field that can be applied to future study designs evaluating the hybrid roles of AI and humans in educational contexts.

Acknowledgements

This work was made possible by NewSchools Venture Fund with funding support from the Gates Foundation. We would also like to acknowledge members of The Learning Agency who conducted or supported the AI benchmarking evaluation of the generative AI-powered tools examined in this study: L Burleigh, Joon Suh Choi, Jules King, and Meg Benner.

Declaration on generative AI

During the preparation of this work, the authors used ChatGPT 5.2 in order to: Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. Laird, M. Dwyer, H. Quay-de la Vallee, Hand in Hand: Schools' Embrace of AI Connected to Increased Risks to Students, Center for Democracy and Technology, 2025. URL: https://cdt.org/wp-content/uploads/2025/10/CDT-2025-Hand-in-Hand-Polling-111225-accessible.pdf.

[2] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, P. Tudorache, Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review, Education Sciences, 13(12) (2023) 1216. doi:10.3390/educsci13121216.

[3] M. Zailani Iman, A. Airlangga Asis, A. Uzma Zein Rahma, Enhancing Personalized Learning: The Impact of Artificial Intelligence in Education, Edu Spectrum: Journal of Multidimensional Education 1(2) (2024) 101–112. doi:10.70063/eduspectrum.v1i2.55.

[4] M. Taşkın, Artificial intelligence in personalized education: Enhancing learning outcomes through adaptive technologies and data-driven insights, Human Computer Interaction 8(1) (2025) 173. doi:10.62802/ygye0506.

[5] D. R. E. Cotton, P. A. Cotton, J. R. Shipway, Chatting and cheating: Ensuring academic integrity in the era of ChatGPT, Innovations in Education and Teaching International 61(2) (2024) 228–239. doi:10.1080/14703297.2023.2190148.

[6] K. Szmyd, E. Mitera, The impact of artificial intelligence on the development of critical thinking skills in students, European Research Studies Journal XXVII(2) (2024) 1022-1039. doi:10.35808/ersj/3876.

[7] W. Strielkowski, V. Grebennikova, A. Lisovskiy, G. Rakhimova, T. Vasileva, AI-driven adaptive learning for sustainable educational transformation, Sustainable Development 33 (2024) 1921-1947. doi:10.1002/sd.3221.

[8] H. Li, AI in education: Bridging the divide or widening the gap? Exploring equity, opportunities, and challenges in the digital age, Advances in Education, Humanities and Social Science Research 8(1) (2023) 355. doi:10.56028/aehssr.8.1.355.2023.

[9] A. Nickow, P. Oreopoulos, V. Quan, The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence (NBER Working Paper No. 27476), National Bureau of Economic Research (2020). doi:10.3386/w27476.
[10] W. Ma, O. O. Adesope, J. C. Nesbit, Q. Liu, Intelligent tutoring systems and learning outcomes: A meta-analysis, Journal of Educational Psychology 106(4) (2014) 901–918. doi:10.1037/a0037123.
[11] A. Létourneau, M. Deslandes Martineau, P. Charland, J. A. Karran, J. Boasen, P. M. Léger, A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education, NPJ science of learning 10(1) (2025) 29. doi:10.1038/s41539-025-00320-7.
[12] A. Gurung, J. Lin, J. Gutterman, D. Thomas, A. Houk, S. Gupta, E. Brunskill, L. Branstetter, V. Aleven, K. Koedinger, Human tutoring improves the impact of AI tutor use on learning outcomes, in: A.I. Cristea et al. (Eds.), Artificial Intelligence in Education, AIED 2025, Springer, Cham, 2025, pp. 393-407. doi:10.1007/978-3-031-98459-4_28.
[13] C. D. Robinson, S. Loeb, High-impact tutoring: State of the research and priorities for future learning (EdWorkingPaper), Retrieved from Annenberg Institute at Brown University, 2021. doi:10.26300/qf76-rj21.
[14] D. L. Roorda, H. M. Y. Koomen, J. L. Spilt, F. J. Oort, The influence of affective teacher-student relationships on students' school engagement and achievement: A meta-analytic approach, Review of Educational Research 81(4) (2011) 493-529. doi:10.3102/0034654311421793.
[15] H. A. Davis, Exploring the contexts of relationship quality between middle school students and teachers, The Elementary School Journal: Special Issue on the Interpersonal Contexts of Motivation and Learning 106 (2006) 193–223. doi:10.1086/501483.
[16] R. C. Pianta, B. K. Hamre, M. W. Stuhlman, Relationships between teachers and children, in: W. M. Reynolds, G. E. Miller (Eds.), Handbook of psychology: Educational psychology, Vol. 7, John Wiley & Sons, Inc., 2003, pp. 199-234. doi:10.1002/0471264385.wei0710.
[17] S. H. Birch, G. W. Ladd, The teacher–child relationship and children's early school adjustment, Journal of School Psychology 35 (1997) 61-79. doi:10.1016/S0022-4405(96)00029-5.
[18] G. Roth, A. Assor, Y. Kanat-Maymon, H. Kaplan, Autonomous motivation for teaching: How self-determined teaching may lead to self-determined learning, Journal of Educational Psychology 99 (2007) 761–774. doi:10.1037/0022-0663.99.4.761.
[19] C. Starke, A. Ventura, C. Bersch, M. Cha, C. de Vreese, P. Doebler, M. Dong, N. C. Krämer, M. Leib, J. Peter, L. Schäfer, I. Soraperra, J. Szczuka, E. Tuchtfeld, R. Wald, N. Köbis, Risks and protective measures for synthetic relationships, Nature Human Behaviour 8(10) (2024) 1834–1836. doi:10.1038/s41562-024-02005-4.
[20] C. E. Tandar, S. S. Bajaj, F. C. Stanford, Social Media and Artificial Intelligence—Understanding Medical Misinformation Through Snapchat's New Artificial Intelligence Chatbot, Mayo Clinic Proceedings: Digital Health 2(2) (2024) 252-254.