Human-AI Cooperation to Improve Tutoring in the United States

Researchers:

Dora Demszky

Susanna Loeb

ana ribeiro

Carly Robinson

Rose E. Wang

Sector(s): Education, Artificial Intelligence (AI)

Fieldwork: FEV Tutor

Location: Southern School District

Sample: 783 tutors and 1,013 K-12 students from Title I schools

Target group: Primary schools Secondary schools Students Teachers Youth

Outcome of interest: Dropout and graduation Student learning Technology adoption

Intervention type: Coaching and mentoring Digital and mobile Computer-assisted learning Online learning Pedagogical innovation Student motivation

Human-computer collaborations could be a powerful way to make use of the strengths of both humans and technology to improve education. In collaboration with FEV Tutor, researchers conducted a randomized evaluation to study whether Tutor CoPilot, an AI language model tool, could enhance tutoring quality and improve student learning outcomes in math. Researchers found that tutor access to Tutor CoPilot improved student learning outcomes measured by student's mastery of topics and that it was especially beneficial for students with lower-rated or less-experienced tutors.

Policy issue

Effective teaching requires diverse skills and experience, as students learn best from educators who can draw on their experience and knowledge of curricula, understand individual needs and can adapt instruction accordingly. However, training new teachers takes time and resources, and traditional training programs often do not match practical requirements. As a result, many new educators lack expert guidance and are left to learning on the job. High-quality teaching is especially important in tutoring, which is one of the most effective ways to improve student outcomes and close educational gaps. Tutoring plays a particularly important role in math, a subject that is difficult to teach yet strongly tied to college completion and future income.

Human-computer collaborations could be a powerful way to make use of the strengths of both humans and technology to improve education. Generative AI, particularly language models, could support new educators with their teaching. However, standard language models have limitations, since they often lack context, give direct answers instead of encouraging critical thinking, and overall have not yet shown clear improvements in learning outcomes. Can tailored language models trained with think-aloud data by experienced educators be designed to support new tutors, enhance tutoring quality, and improve student learning outcomes?

Context of the evaluation

The evaluation took place in a large Southern school district in the United States. The district served over 30,000 students, with the majority from economically disadvantaged and minority backgrounds. Under state policy, students who performed below grade level on the previous spring's state test were eligible for tutoring and inclusion in the study. The study was conducted across grades 3–6 and of those students eligible, eighty percent identified as Hispanic, and two-thirds were considered economically disadvantaged. The study focused on nine schools which received federal funding to support students from low-income families and 1,013 students were identified as eligible for sponsored math tutoring and attended at least one tutoring session after the launch of Tutor CoPilot. The tutoring took place through an in-school, virtual program focused on mathematics.

Researchers partnered with FEV Tutor, a virtual tutoring provider. The study included 783 full-time tutors who worked in the participating students' schools.



Shutterstock.com

Details of the intervention

In collaboration with a Southern school district and FEV Tutor, researchers conducted a randomized evaluation to study whether a tailored language model could enhance tutoring quality and improve student learning outcomes.

The researchers introduced Tutor CoPilot, an AI program designed to improve education outcomes by providing tutors with real-time expert-like guidance during live sessions. Tutor CoPilot was integrated into a virtual tutoring platform with features like a problem display, shared whiteboard, and chat window. When the tutor activated it, Tutor CoPilot analyzed the ongoing content and interaction in the chat and offered tailored suggestions for responses to the tutors. Tutors were able to customize the guidance by editing or regenerating suggestions, or choosing from different strategies such as providing the solution, offering an example, providing a minor correction, giving a similar problem, simplifying the question, or encouraging the student.

The evaluation took place over two months from March to May 2024. The researchers randomly assigned 783 full-time tutors into two groups:

1. Tutor CoPilot group (388 tutors): Tutors were given access to Tutor CoPilot.

2. Comparison group (395 tutors): Tutors were not given access to Tutor CoPilot.

On average the tutors had about 21.5 months tutoring experience. To rate the quality of tutors, the provider observed a random selection of the tutor's sessions and averaged the scores from an observation form. The tutors in the Tutor Co-Pilot Group received training on how to use Tutor CoPilot prior to their first tutoring sessions. Session outcomes were evaluated based on student's mastery of the lesson content required for progressing to the next lesson and measured in "exit tickets." The researchers also collected session chat, whiteboard activity, and overall Tutor Co-Pilot usage. In addition, qualitative interviews were conducted with tutors after the sessions.

The following steps were taken to protect the users' privacy and information in the implementation of the Tutor CoPilot system: The names of the students and tutors were de-identified before any data was shared through external application interfaces, and the conversation used as context was limited to the most recent 10 messages in order to limit the amount of user information shared.

Results and policy lessons

Researchers found that Tutor CoPilot improved student learning outcomes as measured by student mastery of a lesson topic, measured in "exit tickets," was especially beneficial to students of lower-rated and less-experienced tutors, and enhanced the quality of tutor engagement.

Student learning: Students whose tutor had access to Tutor CoPilot had improved learning gains right after the tutoring session. They were four percentage points more likely to pass their exit tickets (66 percent of students in the Tutor CoPilot group compared to 62 percent in the comparison group). The study was not set up to assess Tutor CoPilot's direct effects on statewide end-of-year math test scores.

Impact on tutor effectiveness: Students of lower rated tutors were nine percentage points more likely to pass their exit tickets (65 percent in the Tutor CoPilot group compared to 56 percent in the comparison group). Students of less-experienced tutors were seven percentage points more likely to pass their exit tickets (68 percent in the Tutor CoPilot group compared to 61 percent in the comparison group). The tool helped less-effective tutors achieve outcomes similar to their higher-performing peers in the comparison group.

Tutoring quality: Tutors with access to Tutor CoPilot were more likely than tutors without access to Tutor CoPilot to apply strategies that promoted student understanding, engagement, and skill development, e.g., asking students to further explain or asking questions to guide student thinking. In contrast, tutors without the tool tended to provide more direct or passive support, e.g., giving solution strategies or generic encouragement, compared to tutors with access to the tool.

Perception of Tutor CoPilot by tutors: In post-lesson interviews, tutors generally found Tutor CoPilot helpful, especially for its clear explanations and ability to break down complex concepts in real time. However, some also noted areas for improvement, such as aligning suggestions more closely with students' grade levels. Some responses were described as too advanced, requiring tutors to adapt and simplify the guidance.

Cost-analysis: The total application programming interface cost for integrating the Tutor CoPilot program for 388 tutors over the 2-month study was US$1,419.66, resulting in an estimated annual cost of about US$20 per tutor. This does not include other costs for the tutoring program like salaries.

Use of Results: The tutoring provider FEV Tutor incorporated Tutor CoPilot into all of its tutoring sessions following this study. Although FEV Tutor is no longer active as a company as of January 2025, other companies that offer tutoring services, such as Eedi, have adapted the concept of Tutor CoPilot and implemented it to support human tutors in chat-based tutoring systems.

Taken together, the research on Tutor CoPilot demonstrated that AI-generated guidance, based on expert thinking, can improve tutor quality at a low cost with the potential to scale. As policymakers and researchers continue to explore the future of human-AI collaboration in real-world contexts like education, areas for further research include how well tutors retain the skills gained through real-time AI guidance; expanding Tutor CoPilot to new subjects, age groups, and skill areas; or examining how tutors personalize AI suggestions.

Wang, Rose E., Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. “Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise”. Working Paper, November 2025. https://edworkingpapers.com/sites/default/files/ai24_1054_v2.pdf.

1.