❋ Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise

Rose E. Wang, **Ana T. Ribeiro**, **Carly D. Robinson**, **Susanna Loeb**, and **Dora Demszky**

Stanford University
rewang@cs.stanford.edu

Abstract

Generative AI, particularly Language Models (LMs), has the potential to transform real-world domains with societal impact, particularly where access to experts is limited. For example, in education, training novice educators with expert guidance is important for effectiveness but expensive, creating significant barriers to improving education quality at scale. This challenge disproportionately hurts students from under-served communities, who stand to gain the most from high-quality education and are most likely to be taught by inexperienced educators. We introduce Tutor CoPilot, a novel Human-AI approach that leverages a model of expert thinking to provide expert-like guidance to tutors as they tutor. This study presents the **first randomized controlled trial of a Human-AI system in live tutoring**, involving 900 tutors and 1,800 K-12 students from historically under-served communities. Following a preregistered analysis plan, we find that students working on mathematics with tutors randomly assigned to have access to Tutor CoPilot are 4 percentage points (p.p.) more likely to master topics (p<0.01). Notably, students of lower-rated tutors experienced the greatest benefit, improving mastery by 9 p.p. relative to the control group. We find that Tutor CoPilot costs only $20 per-tutor annually, based on the tutors' usage during the study. We analyze 550,000+ messages using classifiers to identify pedagogical strategies, and find that tutors with access to Tutor CoPilot are more likely to use strategies that foster student understanding (e.g., asking guiding questions) and less likely to give away the answer to the student, aligning with high-quality teaching practices. Tutor interviews qualitatively highlight how Tutor CoPilot's guidance helps them to respond to student needs, though tutors flag common issues in Tutor CoPilot, such as generating suggestions that are not grade-level appropriate. Altogether, our study of Tutor CoPilot demonstrates how Human-AI systems can scale expertise in real-world domains, bridge gaps in skills and create a future where high-quality education is accessible to all students.[1]

1 Introduction

Generative AI, including Language Models (LMs), has the potential to transform real-world domains like education, healthcare and law, which rely on a skilled workforce to handle complex tasks. For example, in education, educators (e.g., teachers and tutors) must be trained to synthesize curriculum knowledge and recognize student needs in order to provide high-quality learning experiences (Hill et al., 2005, Shulman, 1986). Traditionally, experts train novices by sharing their intuition and reasoning refined over years of practice (Darling-Hammond et al., 2017, Tofel-Grehl and Feldon, 2013). However, expert-guided training is costly and difficult to scale. Professional development programs for educators can cost $3,300 per teacher or more annually, with national investments reaching tens of billions of dollars (Kelly et al., 2020, Knight, 2012, Kraft et al., 2018). Additionally,

---

[1]Our pre-registration for this randomized controlled trial can be found here: `https://osf.io/8d6ha`.

these programs demand hours outside the novice's teaching time, which many part-time educators cannot meet (Yoon, 2007). Moreover, traditional training programs are often misaligned to the practical needs of novices because they follow rigid curricula (Boyd et al., 2009, Van Veen et al., 2012, Week, 2019). As a result of expense and misalignment, many novice educators do not have access to expert guidance and must develop their skills on the job, leading to missed learning opportunities for students. This dynamic disproportionately harms students from under-served communities who have the most to gain from improved educational experiences (Boyd et al., 2005, Darling-Hammond and Berry, 2006, Lankford et al., 2002).

LMs may be able to provide real-time guidance for novices at scale, but there are several challenges to enabling this. LMs are trained on Web data (e.g., Wikipedia and Reddit) which differ substantially from real-world K-12 interactions, thus out-of-the-box LMs often fail in real-world learning settings (Achiam et al., 2023, Soldaini et al., 2024, Wang and Demszky, 2023). Current techniques to adapt LMs for real-world settings (e.g., fine-tuning or prompt-engineering) struggle to elicit appropriate behaviors because these approaches focus on surface-level language patterns and overlook the *latent* reasoning processes that expert educators have honed through years of practice to guide their decision-making (Singer, 2023, Team et al., 2023, Wang et al., 2024a). LMs also lack real-world knowledge that is important for delivering high-quality learning experiences, such as knowledge of the curriculum, the student's learning style, previous interactions, and future learning objectives. In contrast, human educators possess this contextual knowledge (e.g., Wang et al. (2024a)). Their knowledge may complement traditional LMs to produce more effective approaches to supporting novices at scale.

We introduce Tutor CoPilot, a Human-AI approach to scale expertise by generating real-time, expert-like suggestions to tutors. We leverage prior work by Wang et al. (2024a) that uses think-aloud protocols to extract the latent expert reasoning from experienced educators and adapts LMs to generate expert-like suggestions. Tutor CoPilot aims to improve the quality of K-12 education at scale by generating actionable guidance that tutors can immediately apply as they tutor, addressing the needs of the tutor in real time and improving the student's live learning experience. This preregistered study presents the **first randomized controlled trial of a Human-AI system in live tutoring**. We conduct an intervention with 900 tutors and 1,800 K-12 students from Title I schools[2] in collaboration with FEV Tutor (a virtual tutoring provider) and a U.S. Southern school district through an in-school, virtual tutoring program for mathematics. In this report, we aim to answer four research questions (RQs):

1. To what extent does Tutor CoPilot affect student learning?

2. Does the effect of Tutor CoPilot on student learning differ by the initial effectiveness of tutors?

3. How does Tutor CoPilot change tutoring quality as measured by the language tutors use with students?

4. How do tutors perceive Tutor CoPilot?

Our findings reveal that **Tutor CoPilot significantly improves student learning outcomes (RQ1)**: Students whose tutors have access to Tutor CoPilot are 4 percentage points (p.p.) more likely to master lesson topics, determined by an intent-to-treat analysis. **Tutor CoPilot particularly benefits lower-rated and less-experienced tutors (RQ2)**, with these tutors improving their students' mastery by up to 9 percentage points over the control, according to a heterogeneity analysis based on tutor ratings from the tutoring provider. **Tutors with access to Tutor CoPilot are more likely to use**

---

[2]Title I schools receive federal funding to support students from low-income families.

high-quality strategies that foster student understanding (RQ3), determined by classifiers that identify high- and low-quality pedagogical strategies. Finally, in our interviews with tutors, **tutors reported that Tutor CoPilot was helpful but indicated room for improvement in its guidance**, such as by generating appropriate grade-level language. With an estimated annual cost of just $20 per tutor based on usage patterns, Tutor CoPilot offers a scalable and cost-effective alternative to traditional, resource-intensive training programs. Overall, our findings demonstrate that Tutor CoPilot is a promising Human-AI approach that combines LMs with task-specific expertise for real-world impact. Tutor CoPilot not only raises the quality of educational experiences for students from under-served communities, but also, more broadly, scales expertise to positively transform critical real-world domains.

2 Related Work

2.1 Training Novices for Complex Real-World Tasks

Real-world domains like education and healthcare require practitioners to draw on a range of skills and navigate unpredictable situations to respond effectively. Teaching, for example, is inherently complex because it involves a broad spectrum of knowledge, including curriculum content, state standards, assessment criteria, practical experience, and an understanding of individual students (Ball et al., 2008, Borko et al., 1992, Lampert, 2001, Willingham, 2008). High-quality teaching is especially critical in interventions like tutoring, which is one of the most effective ways to improve student outcomes and reduce educational disparities (Dietrichson et al., 2017, Nickow et al., 2024), particularly in mathematics—a subject strongly linked to college graduation rates and future earnings (Allensworth and Easton, 2005, Dougherty, 2003, Duncan et al., 2007, Murnane et al., 2000, Watts, 2020). However, training novices is difficult to scale. Professional development programs (e.g., workshops) are often expensive (Heinrich et al., 2014, Kelly et al., 2020, Knight, 2012, Kraft et al., 2018), time-consuming and burdensome (Hill et al., 2020, York et al., 2019). Even if educators do receive training through professional development programs, this form of training is disconnected from the real-time needs of educators because it happens outside of the actual teaching environment and relies on static materials (Hill et al., 2020). Additionally, math is particularly hard to teach because math requires teachers to apply flexible pedagogical strategies to engage students in critical thinking (Hattie, 2008, Wood et al., 1976). Novices often struggle to remediate student mistakes in real-time, missing critical learning opportunities. Our work addresses these challenges by providing a scalable, cost-effective solution that directly provides expert-like guidance to tutors in their live teaching contexts, helping them navigate the complexities of teaching mathematics.

2.2 AI and K-12 Education

Recent advances in AI have sparked excitement about the potential for LMs to transform K-12 education, such as in automated tutoring and feedback generation (Graesser et al., 2004, Khan Academy, 2023, OpenAI, 2023). However, adapting LMs for real-world domains is challenging. LMs are trained on the Web (e.g., Wikipedia) that is vastly different from real-world interactions (e.g., K-12 interactions), and current methods like fine-tuning and prompt engineering struggle to close this gap (Frieder et al., 2023, Ji et al., 2023, Singer, 2023, Wang and Demszky, 2023). These methods rely on surface-level data and explicit rules (Beurer-Kellner et al., 2023, Jurenka et al., 2024, McCoy et al., 2023), which fail to capture the latent expertise—the reasoning honed through years of practice—required for complex real-world tasks like teaching (Blasi, 1995, Polanyi, 1966, Seamster et al., 1993). As a result, LMs often generate bad pedagogical responses, such as giving away answers rather

than fostering critical thinking (Frieder et al., 2023, Singer, 2023, Wang and Demszky, 2023), and these shortcomings have negatively impacted students' educational outcomes (Bastani et al., 2024, Nie et al., 2024). We build on prior work with Bridge (Wang et al., 2024a), an adaptation method that captures expert decision-making by transforming raw think-aloud data with experienced educators into effective instructions for LMs. Bridge has been validated to outperform fine-tuning and prompt-engineering baselines in generating expert-like pedagogical responses.

2.3   Human-AI Collaborative Systems

Human-computer collaborative systems combine the complementary strengths of humans and computers to enhance performance across various domains like healthcare (Lai et al., 2021), writing (Coenen et al., 2021, Lee et al., 2022) and data annotation (Jiang et al., 2021, Kim et al., 2024). Prior research has explored how AI can aid human decision-making on discrete tasks (e.g., selecting treatment for patients) (Cai et al., 2019, Greenes, 2014, Mosquera-Lopez et al., 2014), however researchers have done less work on AI-assistance for tasks that involve continuous interactions. For example, teaching involves dynamic, real-time interactions with students where educators must draw on their knowledge of curricula, assessment standards, and students built through past experiences (Cohen et al., 2003, Hill and Charalambous, 2012, Pianta, 2016, Robinson, 2022, Roorda et al., 2011, Stein, 2007). This grounded, context-specific knowledge is something educators bring, drawing from their real-world experiences. While AI lacks this real-world understanding, it can still provide valuable real-time guidance to support educators' decisions during instructional moments. AI has already shown promise in more scripted environments like call centers, where novice workers benefit from immediate, context-specific suggestions to improve customer interactions (Brynjolfsson et al., 2023). Building on prior work with Bridge (Wang et al., 2024a), we emphasize the importance of the human's role in selecting expert-curated strategies to instruct LMs to generate effective pedagogical suggestions. Prior findings suggest that without this human selection, LMs tend to default to repetitive strategies, whereas experienced educators employ a richer and more diverse set of strategies in their teaching (Wang et al., 2024a). This observation underscores the critical role of the human in guiding the AI to provide the most contextually appropriate and impactful instructional support.

3   Tutor CoPilot

This section describes how Tutor CoPilot works on the *front end* (what the tutor sees) and the *back end* (what our system does under the hood). Figure 1 shows a high-level overview of Tutor CoPilot. The main author of this study worked part-time with FEV Tutor, the tutoring provider, and collaborated closely with several teams (e.g., engineering, design, tutoring operations and curriculum teams) to build Tutor CoPilot and framework for LM-based interventions. Appendix A provides information on how we tested the Tutor CoPilot system prior to deployment.

Integrating into Real-Time Context (Figure 1a).   Tutor CoPilot is directly integrated into tutoring sessions to support live interactions. The current setup embeds Tutor CoPilot within a virtual tutoring platform that features a problem display, shared whiteboard, and chat window. A new Tutor CoPilot button ✳ is added, allowing tutors to easily activate it during sessions for immediate assistance that seamlessly fits into their workflow.

Ensuring User Safety and Privacy (Figure 1b).   Once activated, Tutor CoPilot pulls relevant information from the ongoing interaction, such as the conversation context, lesson topic and the
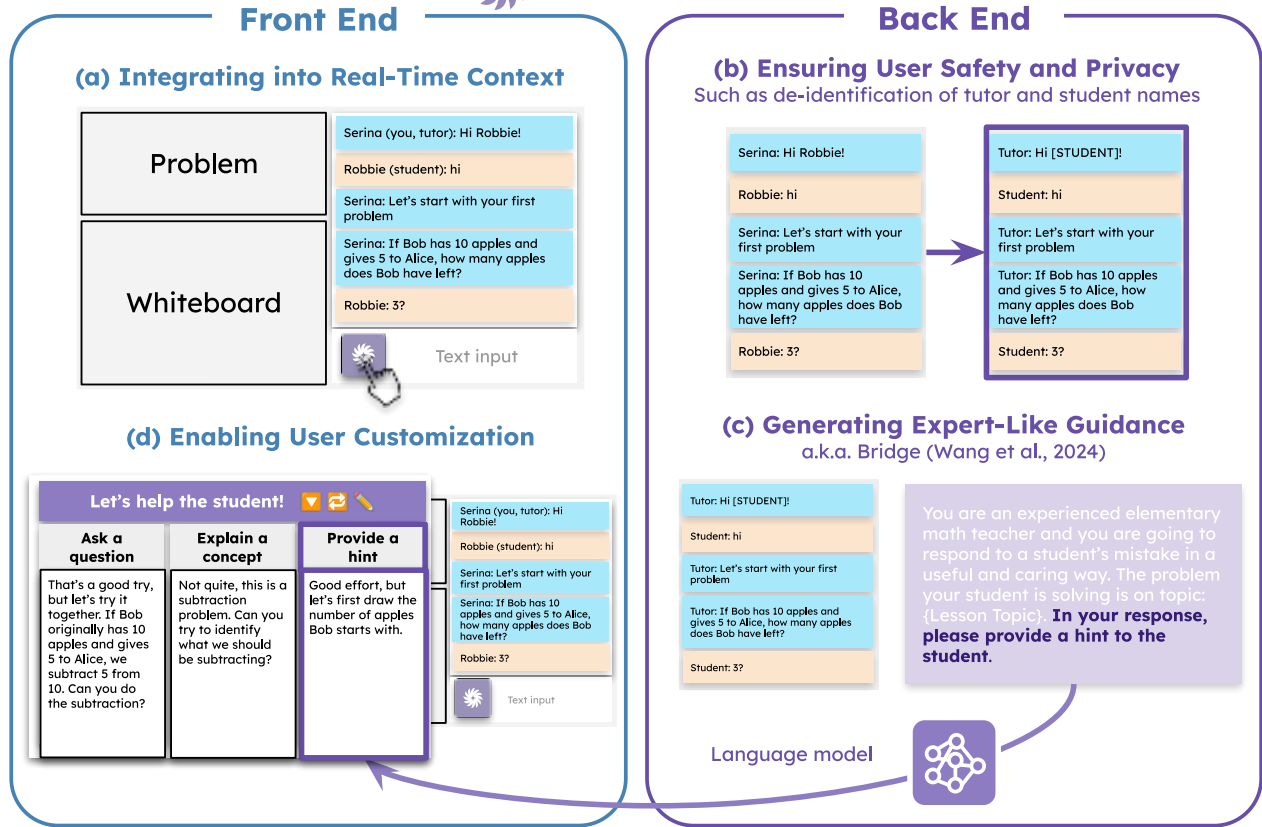
Figure 1: **Illustration of Tutor CoPilot.** (a) *Tutor CoPilot is integrated into live contexts* as a button which the tutor can activate for real-time assistance during their tutoring sessions. (b) *Tutor CoPilot applies user safety and privacy practices*, such as automatically de-identifying student and tutor names and limiting the amount of user information sent to external LM services. (c) *Tutor CoPilot generates expert-like guidance* by leveraging the Bridge method (Wang et al., 2024a) which captures expert decision-making from their verbalized reasoning patterns. (d) *Tutor CoPilot enables user customization*. The tutor can customize the guidance by editing (✏️), re-generating (🔁), or selecting a different strategy (🔽).

requested strategy.[3] Currently, the conversation context is based on chat interactions, but the system is flexible and can be adapted for in-person tutoring by processing speech or visual inputs, like whiteboard activity. To ensure user safety and privacy, we automatically de-identify student and tutor names retrieved from the roster database using placeholders "[STUDENT]" and "[TUTOR]" via Edu-ConvoKit (Wang and Demszky, 2024). Additionally, when using external LM services (e.g., OpenAI), we limit shared conversation context to the 10 most recent messages to minimize data exposure. These measures lay the groundwork for future privacy safeguards with human-AI systems.

---

[3]The strategy is chosen either from default options or a dropdown menu (Figure 1d).

Generating Expert-Like Guidance (Figure 1c).   Leveraging Bridge (Wang et al., 2024a), we generate expert-like suggestions based on the de-identified conversation, lesson topic, and the chosen strategy (e.g., "provide a hint"). This approach *lightens the cognitive load* on tutors by eliminating the need for them to figure out how to prompt the model themselves and allows them to focus on delivering high-quality instruction. Bridge is also open-sourced for others to use.

Enabling User Customization (Figure 1d).   Tutors can personalize the generated guidance by editing suggestions (✏️), regenerating them (🔄), or selecting a different strategy (🔽). Available strategies include providing a solution, a worked example, a minor correction, a similar problem, simplifying the question, affirming the correct answer, and encouraging the student. Selecting a new strategy updates the suggestion in the response box. Unlike typical autocompletion systems that provide a single response, Tutor CoPilot presents multiple suggestions based on different strategies. Our approach gives tutors a range of options to maintain quality while preserving their autonomy in effectively addressing student needs.

4   Study Design

We conducted a randomized controlled trial to evaluate whether Tutor CoPilot, a human-LM system, can scale real-time expertise and improve tutoring quality. Our theory of change centers on the observation that many tutors lack the experience to provide effective responses in the moment. As a result, they generate low-quality responses, such as giving generic feedback (e.g., "Good job") or giving away the answer, which can hinder student learning. By offering real-time guidance tailored to the conversation and following high-quality strategies, Tutor CoPilot empowers tutors with the right language to develop the student's understanding. Thus, the tool builds tutors' skills while directly improving learning interactions.

The study started at the end of March 2024 and lasted for two months. We preregistered our primary hypotheses and analysis plan on the Open Science Framework platform prior to accessing the data which can be found: https://osf.io/8d6ha.

4.1   Participants and Randomization

We partnered with FEV Tutor (a virtual tutoring provider) and a large southern U.S. school district. The school district serves more than 30,000 students, with the majority from ethnic/racial minority groups and economically disadvantaged backgrounds. Nine schools from the school district participated in our study. The district identified 1787 students across these schools to participate. In accordance with state policy for administering accelerated services, students were eligible for tutoring services and for the study if they had performed below grade level on the state test the previous spring. The study was conducted with with elementary and middle school students in grades 3-8. The majority (80%) of our sample identifies as Hispanic, and 67% of the sample is classified as economically disadvantaged. The full sample description is provided in Appendix G.

The study also involved 874 full-time tutors assigned to work in those students' schools. We conducted randomization at the tutor level, randomly assigning tutors to one of two conditions:

- Treatment group: Tutors get access to Tutor CoPilot in their tutoring sessions.
- Control group: Tutors do not have access, and tutor as normal in their tutoring sessions.

Our initial assignment involved 900 tutors, evenly divided between treatment and control (treatment = 450, control = 450). However, due to attrition between the initial assignment time and the study

launch, our study sample size at the launch was 782 (treatment = 386, control = 396). Balance checks show that the control and treatment groups do not differ statistically on any baseline variables; these results are detailed in Appendix B. Appendix D details the training tutors received prior to the study launch; treatment tutors were trained on how to use the tool to minimize novelty effects, while control tutors received a standard training module from the provider's existing materials.

4.2 Data

Our study uses three types of data about the tutor, student and session. Appendix E provides more details on our study's data.

Tutors. For tutors, we collected their treatment group assignment, gender, pre-study quality rating, and pre-study tutoring experience. The quality rating was derived from the provider's manual observations of a random sample of the tutor's sessions, averaged from scores on an observation rubric; it takes on continuous values between -1 and 1. The tutor's tutoring experience is the number of months the tutor had been working with the tutoring provider leading up to the start of the study.

Students. For students, we collected their gender, race, their pre-/post-study NWEA MAP Math and Reading Scores, and other relevant covariates. The NWEA MAP assessments are standardized tests administered three times a year, tracking students' academic growth over time.[4]

Tutoring sessions. We tracked session outcomes such as whether the student passes their exit ticket, participation points, and student post-session survey responses. **Exit tickets are particularly critical** as they assess student mastery of each topic and directly influence the student's progression through the curriculum. Students must pass the exit ticket to advance to the next lesson, meaning students who struggle on exit tickets may encounter delays in their learning progression, limiting their exposure to new material. As a result, the exit ticket pass rate is not just a measure of immediate comprehension—it plays a key role in determining the pace at which students advance through their academic content. Moreover, the number of completed exit tickets by a student is a significant predictive of future MAP test performance: one additional exit ticket completed is associated with a 0.06 point increase on the test; we report this in Appendix F.

Additionally, we collected each session's chat and whiteboard activity. Our final analysis sample includes 4,136 sessions with 550,000+ chat messages. We also collected data on Tutor CoPilot usage. This includes 2,000+ uses of Tutor CoPilot; we define a use of Tutor CoPilot to correspond to either an initial activation of Tutor CoPilot (Figure 1 a) or the tutor clicking on a different strategy from the dropdown options (Figure 1 d). Any reference to "use of Tutor CoPilot" should be understood according to this definition.

5 Methods

This study tested the causal impact of Tutor CoPilot on student learning outcomes and tutor practices. Below, we summarize our methods for analyzing the following research questions.

---

[4]Following our pre-registration, we imputed missing covariates for students and tutors. For categorical variables, we added an additional "missing" category. For continuous variables, we assigned the predicted value based on the other present covariates.

RQ1: To what extent does Tutor CoPilot affect student learning? We employed an intention-to-treat (ITT) regression analysis to determine how offering the Tutor CoPilot to tutors predicts student outcomes.[5] Equation 1 describes our primary model:

$$Y_{ijk} = \alpha + \beta\text{Treatment}_j + X_i\gamma + \omega_k + \epsilon_{ij} \tag{1}$$

where Y_{ijk} is the session-level outcome of interest (e.g., exit ticket passed) for student i working with tutor j in a given class (combination of school and grade) k, Treatment$_j$ is the indicator for whether tutor j is in the treatment group, X_i is a vector of student-level covariates, ω_k is the fixed effect for class (school and grade), and ϵ_{ij} is the residual clustered at the student-tutor pair level. The student covariates include categorical indicators for the student's gender, race, free and reduced lunch, special education, and limited English proficiency, as well as a continuous variable of the student's pre-study MAP math score.

RQ2: Does the effect of Tutor CoPilot on student learning differ by the initial effectiveness of tutors? We performed a heterogeneity analysis using the exit ticket passing rate as the primary outcome from our primary model (Equation 1). Specifically, we categorized the tutor's quality rating and tutoring experience into tercile indicators and interacted these indicators with the student's exit ticket passing rate to examine how the effect of Tutor CoPilot varies based on the tutor's initial effectiveness. This approach allowed us to assess whether more or less effective tutors benefit differently from using Tutor CoPilot in terms of improving student performance.

RQ3: How does Tutor CoPilot change tutoring quality as measured by the language tutors use with students? We investigated whether treatment tutors employed higher-quality instructional strategies compared to control tutors by leveraging NLP methods, in a three-step process. First, we drew on prior work to define a taxonomy of high-quality strategies that foster students' understanding and low-quality strategies that lead students directly to the solution or passively engage students (Boaler, 2013, Carpenter et al., 1999, 2003, Easley and Zwoyer, 1975, Lester, 2007, Loewenberg Ball and Forzani, 2009, Wang et al., 2024a). Table 1 reports the final taxonomy. Second, we trained machine learning classifiers to identify these strategies at scale. We did so by (i) randomly sampling 3,000 tutor messages, blind to study condition, (ii) prompting GPT-4 with our taxonomy of strategies to perform a first pass of labeling these examples, (iii) manually verifying and fixing labels, (iv) fine-tuning RoBERTa models on the labeled data with a class-balancing loss (Cui et al., 2019, Liu, 2019). Our classifiers achieved high performance, with F1-scores ranging between 0.65 and 0.90. Finally, we used these classifiers to identify the strategies in our entire data of 550,000+ messages and compared the likelihood of tutors using these strategies across conditions by computing the z-scored log-odds ratio (Monroe et al., 2008) of each strategy. Appendix I provides more details on this setup.

RQ4: How do tutors perceive Tutor CoPilot? To understand the treatment tutors' experiences and gather feedback on Tutor CoPilot, we conducted interviews with approximately 20 treatment tutors a week after the study concluded. These interviews explored their usage of Tutor CoPilot and how their use evolved over time. We report the qualitative findings and themes that emerged from these interviews. Additional details on the interview setup are available in Appendix J.

---

[5]While the ITT analysis captures the overall impact of the intervention, it ignores whether the tutor used Tutor CoPilot during a session. To disentangle the effect of merely having access to the tool from the effect of actually using it, we extended our analysis using a two-stage least squares (2SLS) regression and report those results in Appendix H.

Quality Category	Strategy Name (Frequency)	Definition	Examples	F1
High	Prompt Student to Explain (2%)	The tutor prompts the student to explain a concept, rule, or their reasoning.	"Go ahead and try to explain how you got the answer."	0.89
High	Ask Question to Guide Thinking (5%)	The tutor asks the student a question to help them think the problem.	"What number can we multiply the number 10 to get an equal value of 100?"	0.90
High	Affirm Student's Correct Attempt (9%)	The tutor affirms the student's correct attempt.	"Yes, 20 is the correct answer."	0.65
Low	Ask Student to Retry (1%)	The tutor asks the student to recheck their work or try again.	"Please recheck your answer."	0.73
Low	Give Away the Answer / Explanation (9%)	The tutor provides the answer or explanation to the student.	"So, the greatest number will be 7520."	0.76
Low	Give Away the Solution Strategy (11%)	The tutor provides a strategy for solving the problem.	"We can order the list according to the hundredths place value."	0.79
Low	Encourage Student in Generic Way (12%)	The tutor encourages the student without being specific about the student's attempt.	"That's a good try!"	0.81

Table 1: Taxonomy of high- and low-quality strategies, including their definitions, examples, and frequency over the labelled dataset. We train binary classifiers to identify these strategies at scale and report their test F1 score as well.

6   Results

We report the results to our research questions below. Appendix G reports descriptive results, such as the descriptive student numbers and tutor adoption of Tutor CoPilot; Appendix L reports the tutor-level analysis; and Appendix C reports notes on compliance.

RQ1: To what extent does Tutor CoPilot affect student learning?   Table 2 reports the estimates of effects of treatment assignment on session-level outcomes. "Exit Ticket Attempted" indicates whether the student attempted the exit ticket during the session, and "Exit Ticket Passed (Conditional)" refers to whether the student conditioned that they attempted the exit ticket. "Exit Ticket Passed (Unconditional)" refers to whether the student passed the exit ticket (i.e., if the student didn't attempt the exit ticket, they did not pass).

We observed there is a ***significant positive treatment effect on students passing their exit tickets***: Students working with treatment tutors were 4 percentage points (p.p.) more likely to pass their exit tickets ($p < 0.01$, $62\% \rightarrow 66\%$ student passing rate from control to treatment).[6] We also found that students were 2 p.p. more likely to attempt the exit ticket ($p < 0.1$), and conditional on their attempt, they were 3 p.p. more likely to pass the exit ticket ($p < 0.01$). We found positive estimates

---

[6]Appendix H reports results on the tutors *using* Tutor CoPilot, rather than just having access to it, where students were **14 p.p.** more likely to pass their exit tickets ($p < 0.01$).

| | Panel A. Session outcomes | | | | |
Dependent Variable:	Participation Points	Participation Points (Standardized)	Exit Ticket Attempted	Exit Ticket Passed (Conditional)	Exit Ticket Passed (Unconditional)
Treatment Effect	0.09	0.01	0.02^{+}	0.03^{*}	0.04^{**}
	(0.27)	(0.03)	(0.01)	(0.01)	(0.01)
Control Mean	14.07	0.016	0.84	0.73	0.62
	(0.20)	(0.02)	(0.01)	(0.01)	(0.01)
n	4136	4136	4136	3521	4136
	Panel B. Student survey outcomes				
Dependent Variable:	My tutor cared understanding math over memorizing the solution.	My tutor cared about how well I do in math.	Even when math is hard, I know I can learn it.	Session Rating	Tutor Rating
Treatment Effect	-0.004	0.02	0.02	-0.002	0.03
	(0.055)	(0.05)	(0.06)	(0.036)	(0.04)
Control Mean	4.19	4.31	4.24	4.77	4.74
	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)
n	1931	1931	1931	1948	1952

Note. $^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Estimates are from our primary model, which controls for baseline math scores, student demographics, and a fixed effect for strata (school × grade). Participation points are standardized within-sample and by grade. The survey items are on a 5-point scale where higher is better.

Table 2: ITT analysis on student session-level outcomes. The parentheses report the standard error.

of the treatment assignment on the participation points and a mix of positive and negative estimates on the student survey outcomes. These estimates are not significant.

RQ2: Does the effect of Tutor CoPilot on student learning differ by the initial effectiveness of tutors? Figure 2 reports the effect of Tutor CoPilot on student learning differ by the initial effectiveness of tutors measured by either the tutor's quality rating or experience. The results indicate *substantial benefits for tutors with lower initial effectiveness* in both categories. In Figure 2a, lower-rated tutors experienced a 9 p.p. increase in student's passing their exit ticket ($56\% \rightarrow 65\%$ student passing rate from control to treatment). In Figure 2a, lower-experienced tutors experienced a 7 p.p. increase ($61\% \rightarrow 68\%$). We saw these treatment effects diminish with increasing tutor effectiveness, but notably, students of lower-rated or less-experienced tutors in the treatment group performed at or above the level of students with higher-rated or more experienced tutors in the control group. This suggests that Tutor CoPilot helped less-effective tutors achieve outcomes comparable to their more-effective peers.

RQ3: How does Tutor CoPilot change tutoring quality as measured by the language tutors use with students? Figure 3 reports the z-scored log-odds of strategies used between treatment and control tutors, showing the impact of Tutor CoPilot on the types of strategies employed during tutoring sessions. We found that *treatment tutors were significantly more likely to use strategies aimed at fostering student skills and comprehension*. For example, "prompting the student to explain" and "asking questions to guide thinking" were used approximately 2 standard deviations more frequently on a log-odds scale in treatment sessions compared to control sessions. These

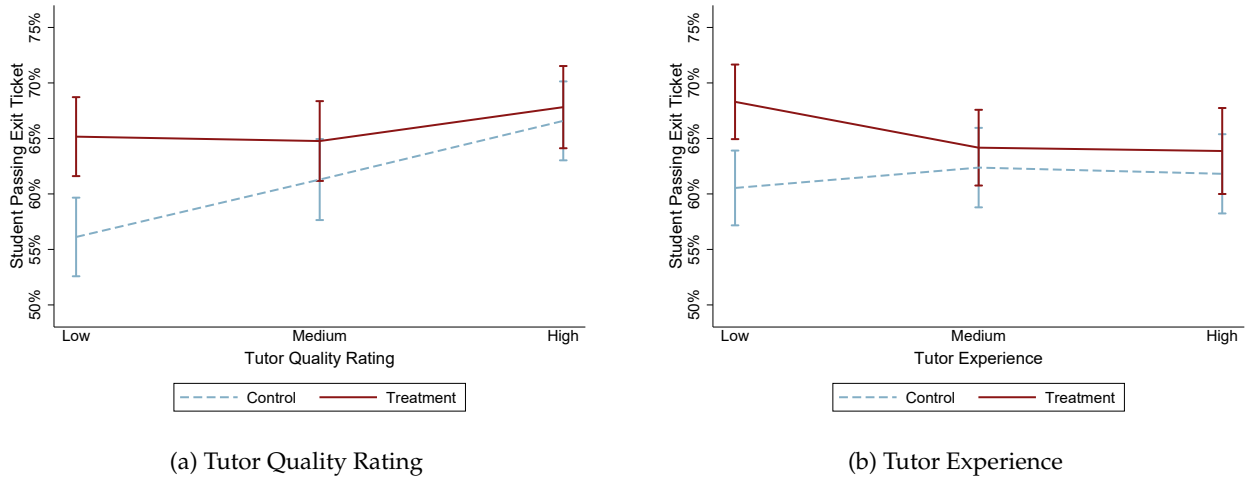(a) Tutor Quality Rating   (b) Tutor Experience

Figure 2: Heterogeneity analysis by tutor initial effectiveness on student learning. (a) reports by the tutor's initial quality rating and (b) by tutor's tutoring experience.
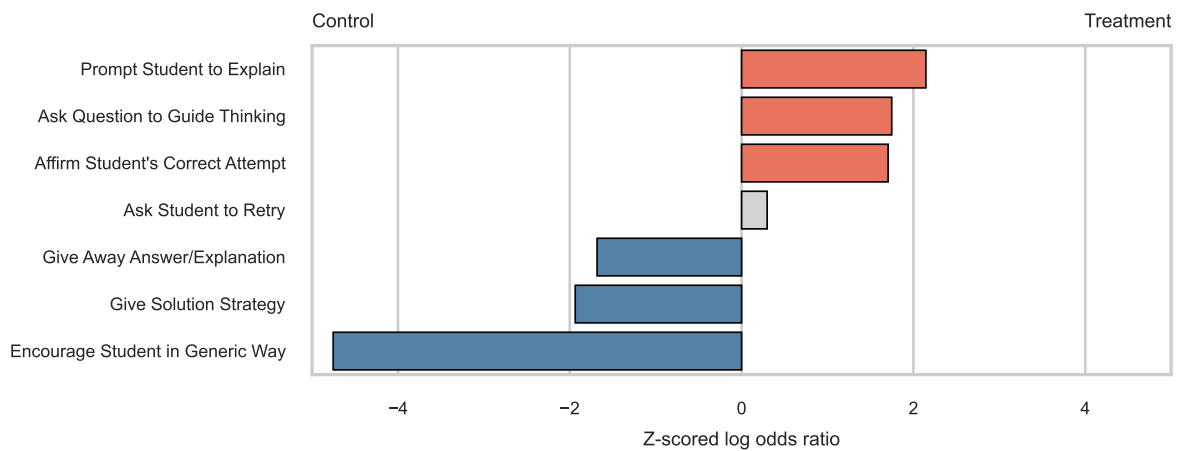

Figure 3: Strategies more likely to be used by control tutors (left) vs. treatment tutors (right). Strategies with a z-score below 1 standard deviation are shaded in gray. Control tutors tended to rely on solution-focused, passive strategies, while treatment tutors more frequently used strategies that promote deeper student engagement and comprehension.

strategies align with expert-recommended practices for promoting deeper learning. In contrast, control tutors tended to rely on strategies that focus on directly leading students to the solution or providing passive support. These strategies included giving away the answer or providing the solution, which help students complete tasks but do less to develop their deeper understanding. This contrast highlights the difference in instructional approaches, with treatment tutors more frequently employing strategies that foster active student engagement and cognitive development. These findings provide evidence that treatment tutors might have achieved better student learning outcomes through the use of more expert-like teaching strategies.

11

RQ4: How do tutors perceive Tutor CoPilot? Tutors generally found Tutor CoPilot helpful, particularly for its ability to provide well-phrased explanations and break down complex concepts on the spot. They highlighted its usefulness in explaining difficult topics, such as differentiating between mean and median, and appreciated the clear definitions and hints it generated. However, some tutors noted areas for improvement, mentioning that the tool's suggestions occasionally lacked alignment with students' grade levels. A common issue was that the responses were sometimes perceived as "too smart," requiring tutors to simplify and adapt them for clarity. More details on the tutor interviews can be found in Appendix J.

7 Limitations

Our study presents promising results on Human-AI approaches for scaling real-time expertise. However, we note the following key limitations to our work:

Generalizability within education. Our intervention on Tutor CoPilot is conducted with novice tutors who teach young students from under-served communities (majority Hispanic) based in a Southern school district in the United States. The learning needs of these students may differ significantly from those of students with different demographics, in other parts of the United States, in other countries, or in more affluent learning environments. Additionally, while we observed substantial improvements in students' proximal learning gains (i.e., the students' exit ticket performance), we did not find statistically significant improvements in end-of-year math test scores; please refer to Appendix K for this analysis. This may be due to the little variation in students' exposure to the treatment and the relatively short duration of the study (two months). Future research would benefit from randomizing students to better assess the impact of the intervention.

Generalizability to other domains. While our study focuses on applications to education, our Human-AI collaborative approach demonstrated by Tutor CoPilot has broader potential across various domains that rely on human expertise, such as law and healthcare. In these fields, professionals also face dynamic, high-stakes decision-making environments where expertise is built through continuous interactions and contextual knowledge. Just as AI-assisted tutoring aids educators in delivering expert-like guidance in real time, similar systems could support novice physicians during patient consultations or novice lawyers in legal research and case preparation. However, the transferability of our findings to these domains will depend on the specific tasks, workflows, and levels of domain complexity involved. Careful adaptation of AI assistance tools is important to meet the demand of these professions and align with human expertise.

Multi-modality. Our study collaborates with a chat-based tutoring platform, which is well-suited for real-time use of language models. Future work would benefit from extending Human-AI approaches to include other modalities—such as vision (e.g., whiteboard or student face) and speech (e.g., student and tutor voice)—to incorporate naturalistic modalities for interaction. Including these modalities would provide a more comprehensive understanding of the student and thus enable better expert-like guidance during the interaction. The shift towards multi-modal approaches also raises interesting challenges on how to provide real-time guidance through different modalities, such as via whiteboard drawing or verbal explanations.

User safety. While our study de-identifies names, user identities can still be inadvertently disclosed through non-name information, such as email addresses, phone numbers, or personal anecdotes shared by students and tutors. Similar privacy concerns arise when incorporating vision and speech

modalities, highlighting the need for better safeguards. Additionally, we observe a trade-off between performance and safety. In our study, to mitigate over-exposure, we limited the conversation context passed to external APIs; in initial pilot studies, we varied the conversation context between 5 to 10 messages and found that tutors preferred the quality of interactions when more context was included. Balancing these factors will be important for future research.

8 Conclusion and Discussion

This study introduces Tutor CoPilot, a human-AI system to scale real-time expertise in K-12 tutoring. It represents the first randomized controlled trial of AI-supported live tutoring. Our findings demonstrate that AI-generated guidance—based on expert thinking—can significantly improve tutoring quality, particularly for less experienced tutors. Notably, Tutor CoPilot improves student learning at a low cost of $20 per tutor annually, which is far cheaper than traditional training programs costing thousands of dollars. We also find that Tutor CoPilot helps tutors adopt high-quality strategies that promote deeper learning. Our approach offers a practical and scalable solution for improving education, especially for under-served communities that stand to benefit the most from better quality education.

Looking ahead, our work can extend in several exciting directions. First, we aim to explore how well novices retain the skills they acquired from real-time expert guidance. This analysis will shed light on the long-term effects of the use of AI systems in real-world settings. Second, expanding Tutor CoPilot to other skill areas—such as developing collaborative learning—and applying it in different subjects, age groups, or system designs will provide insights into its broader impact. Last, we observed significant variation in how tutors adapted or personalized AI-generated suggestions, highlighting interesting human-AI collaborative dynamics that merit further study.

As we continue to explore the future of human-AI collaboration in real-world domains, Tutor CoPilot demonstrates the potential to scale expert-level support and delivering high-quality experiences where they are needed most.

Acknowledgments

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Elaine Allensworth and John Q Easton. The on-track indicator as a predictor of high school graduation, 2005.

Deborah Loewenberg Ball, Mark Hoover Thames, and Geoffrey Phelps. Content knowledge for teaching: What makes it special? 2008.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakcı, and Rei Mariman. Generative ai can harm learning. *Available at SSRN*, 4895486, 2024.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Prompting is programming: A query language for large language models. *Proc. ACM Program. Lang.*, 7(PLDI), jun 2023. doi: 10.1145/3591300. URL https://doi.org/10.1145/3591300.

Gary L Blasi. What lawyers know: Lawyering expertise, cognitive science, and the functions of theory. *J. Legal Educ.*, 45:313, 1995.

Jo Boaler. Ability and mathematics: The mindset revolution that is reshaping education. Forum, 2013.

Hilda Borko, Margaret Eisenhart, Catherine A Brown, Robert G Underhill, Doug Jones, and Patricia C Agard. Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for research in mathematics education*, 23(3):194–222, 1992.

Donald Boyd, Hamilton Lankford, Susanna Loeb, and James Wyckoff. Explaining the short careers of high-achieving teachers in schools with low-performing students. *American economic review*, 95 (2):166–171, 2005.

Donald J Boyd, Pamela L Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. Teacher preparation and student achievement. *Educational evaluation and policy analysis*, 31(4):416–440, 2009.

Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. Technical report, National Bureau of Economic Research, 2023.

Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

Thomas P Carpenter, Elizabeth Fennema, M Loef Franke, Linda Levi, and Susan B Empson. Children's mathematics. *Cognitively Guided*, 8, 1999.

Thomas P Carpenter, Megan Loef Franke, and Linda Levi. *Thinking mathematically*. Portsmouth, NH: Heinemann, 2003.

Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. Wordcraft: A human-ai collaborative editor for story writing. *arXiv preprint arXiv:2107.07430*, 2021.

David K Cohen, Stephen W Raudenbush, and Deborah Loewenberg Ball. Resources, instruction, and research. *Educational evaluation and policy analysis*, 25(2):119–142, 2003.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

Linda Darling-Hammond and Barnett Berry. Highly qualified teachers for all. *Educational leadership*, 64(3):14, 2006.

Linda Darling-Hammond, Maria E Hyler, and Madelyn Gardner. Effective teacher professional development. *Learning policy institute*, 2017.

Jens Dietrichson, Martin Bøg, Trine Filges, and Anne-Marie Klint Jørgensen. Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of educational research*, 87(2):243–282, 2017.

Christopher Dougherty. Numeracy, literacy and earnings: Evidence from the national longitudinal survey of youth. *Economics of education review*, 22(5):511–521, 2003.

Greg J Duncan, Chantelle J Dowsett, Amy Claessens, Katherine Magnuson, Aletha C Huston, Pamela Klebanov, Linda S Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, et al. School readiness and later achievement. *Developmental psychology*, 43(6):1428, 2007.

J Al Easley and Russell E Zwoyer. Teaching by listening-toward a new day in math classes. *Contemporary Education*, 47(1):19, 1975.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.

Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36:180–192, 2004.

Robert Greenes. *Clinical decision support: the road to broad adoption*. Academic Press, 2014.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

John Hattie. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge, 2008.

Carolyn J Heinrich, Patricia Burch, Annalee Good, Rudy Acosta, Huiping Cheng, Marcus Dillender, Christi Kirshbaum, Hiren Nisar, and Mary Stewart. Improving the implementation and effectiveness of out-of-school-time tutoring. *Journal of Policy Analysis and Management*, 33(2):471–494, 2014.

Heather C Hill and Charalambos Y Charalambous. Teacher knowledge, curriculum materials, and quality of instruction: Lessons learned and open issues. *Journal of Curriculum Studies*, 44(4):559–576, 2012.

Heather C Hill, Brian Rowan, and Deborah Loewenberg Ball. Effects of teachers' mathematical knowledge for teaching on student achievement. *American educational research journal*, 42(2): 371–406, 2005.

Heather C Hill, Kathleen Lynch, Kathryn E Gonzalez, and Cynthia Pollard. Professional development that improves stem outcomes. *Phi Delta Kappan*, 101(5):50–56, 2020.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. Supporting serendipity: Opportunities and challenges for human-ai collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, et al. Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*, 2024.

Sean Kelly, Robert Bringe, Esteban Aucejo, and Jane Cooley Fruehwirth. Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28:62–62, 2020.

Khan Academy. Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access. https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access, 2023. [Online; accessed 4-June-2024].

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*, 2024.

David S Knight. Assessing the cost of instructional coaching. *Journal of Education Finance*, pages 52–80, 2012.

M. A. Kraft, D. Blazar, and D. Hogan. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4):547–588, 2018. URL https://doi.org/10.3102/0034654318759268.

Yi Lai, Atreyi Kankanhalli, and Desmond Ong. Human-ai collaboration in healthcare: A review and research agenda. 2021.

M Lampert. Teaching problems and the problems of teaching. *Yale University*, 2001.

Hamilton Lankford, Susanna Loeb, and James Wyckoff. Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational evaluation and policy analysis*, 24(1):37–62, 2002.

Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022.

Frank K Lester. *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics*. IAP, 2007.

Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Deborah Loewenberg Ball and Francesca M Forzani. The work of teaching and the challenge for teacher education. *Journal of teacher education*, 60(5):497–511, 2009.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

Clara Mosquera-Lopez, Sos Agaian, Alejandro Velez-Hoyos, and Ian Thompson. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. *IEEE reviews in biomedical engineering*, 8:98–113, 2014.

Richard J Murnane, John B Willett, Yves Duhaldeborde, and John H Tyler. How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, 19(4):547–568, 2000.

Andre Nickow, Philip Oreopoulos, and Vincent Quan. The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. 2024.

Allen Nie, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. The gpt surprise: Offering large language model chat in a massive coding class reduced engagement but increased adopters' exam performances. Technical report, Center for Open Science, 2024.

OpenAI. Gpt-4 technical report, 2023.

Robert C Pianta. Teacher–student interactions: Measurement, impacts, improvement, and policy. *Policy insights from the behavioral and brain sciences*, 3(1):98–105, 2016.

Michael Polanyi. The logic of tacit inference. *Philosophy*, 41(155):1–18, 1966.

Carly D Robinson. A framework for motivating teacher-student relationships. *Educational Psychology Review*, 34(4):2061–2094, 2022.

Debora L Roorda, Helma MY Koomen, Jantine L Spilt, and Frans 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):493–529, 2011.

Thomas L Seamster, Richard E Redding, John R Cannon, Joan M Ryder, and Janine A Purcell. Cognitive task analysis of expertise in air traffic control. *The international journal of aviation psychology*, 3(4):257–283, 1993.

Lee S Shulman. Those who understand: Knowledge growth in teaching. *Educational researcher*, 15(2):4–14, 1986.

Natasha Singer. In classrooms, teachers put a.i. tutoring bots to the test, Jun 2023. URL https://www.nytimes.com/2023/06/26/technology/newark-schools-khan-tutoring-ai.html.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

MK Stein. How curriculum influences student learning. *Second handbook of research on mathematics teaching and learning/Information Age*, 2007.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Colby Tofel-Grehl and David F Feldon. Cognitive task analysis–based training: a meta-analysis of studies. *Journal of cognitive engineering and decision making*, 7(3):293–304, 2013.

Klaas Van Veen, Rosanne Zwart, and Jacobiene Meirink. What makes teacher professional development effective?: A literature review. *Teacher learning that matters*, pages 3–21, 2012.

Rose Wang and Dorottya Demszky. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *18th Workshop on Innovative Use of NLP for Building Educational Applications*, June 2023.

Rose Wang and Dorottya Demszky. Edu-convokit: An open-source library for education conversation data. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 61–69, 2024.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.120. URL https://aclanthology.org/2024.naacl-long.120.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024b.

Tyler W Watts. Academic achievement and economic attainment: Reexamining associations between test scores and long-run earnings. *Aera Open*, 6(2):2332858420928985, 2020.

Education Week. Blind Spots in Teacher Professional Development. https://www.edweek.org/leadership/blind-spots-in-teacher-professional-development, 2019.

Daniel T Willingham. Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, 109 (4):21–32, 2008.

David Wood, Jerome S. Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976.

Kwang Suk Yoon. Reviewing the evidence on how teacher professional development affects student achievement. 2007.

Benjamin N York, Susanna Loeb, and Christopher Doss. One step at a time: The effects of an early literacy text-messaging program for parents of preschoolers. *Journal of Human Resources*, 54(3): 537–566, 2019.

A   Pilot Studies of Tutor CoPilot

Before we fully launched Tutor CoPilot, we conducted two pilot studies to ensure its useability and performed data pulls to ensure our databases accurately tracked information needed for our analysis. These pilot studies involved 10-20 tutors that were not a part of the main intervention. The pilot studies played a critical role in the success of Tutor CoPilot because they identified system bottlenecks that made Tutor CoPilot difficult to use. One main issue from the first pilot study was the response time of Tutor CoPilot: Data analyses, interviews and surveys with tutors surfaced frustrations that the tool took longer than 30 seconds to respond. We profiled Tutor CoPilot and identified inefficiencies in our data retrieval pipeline. The second pilot study showed significant response time improvements and strong tutor satisfaction. This gave us the green light to deploying the tool at scale for our study.

B   Balance Checks and Attrition

We conducted balance checks to ensure that the tutor sample is comparable across treatment and control groups in terms of key covariates. Table 3 presents balance checks on tutor characteristics, including gender, years of tutoring experience, and pre-study quality ratings. Table 4 reports attrition, defined as tutors who were assigned to a group but did not conduct any tutoring sessions during the study period. After excluding tutors with no sessions, Table 5 presents balance checks for the remaining sample, including an additional check on the amount of tutoring conducted by assignment group. For all balance and attrition checks, we find no statistically significant differences between the treatment and control groups, indicating that random assignment was effective and the sample remains balanced after accounting for attrition.

	Control (N=450)	**Treatment** (N=429)	*p-value*
Gender (Is Female)	0.56	0.52	NS
Experience (Months)	21	22	NS
Quality Rating	0.41	0.42	NS

Note. NS = Not Significant. Experience reports the number of months the tutor has been with the tutoring provider. Quality Rating reports the tutor's tutoring quality rating determined prior to the study by the tutoring provider based on session observations & quality rubric scores.

Table 3: Balance check on tutor sample at the start of the study.

	Control	**Treatment**	*p-value*
Original Sample Size	450	429	
Attrition Rate	11%	10%	NS
Final Sample Size	396	386	

Table 4: Attrition rate on tutor sample. We define attrition as not having any tutoring sessions within our data sample.

	Control (N=396)	**Treatment** (N=386)	*p-value*
Gender (Is Female)	0.55	0.52	NS
Experience (Months)	21	22	NS
Quality Rating	0.41	0.42	NS
# Sessions during Study	5	5	NS
Total Session Time during Study (Minutes)	199	193	NS

Note. NS = Not Significant. Experience reports the number of months the tutor has been with the tutoring provider. Quality Rating reports the tutor's tutoring quality rating determined prior to the study by the tutoring provider based on session observations & quality rubric scores.

Table 5: Balance check on study's actual tutor sample (post attrition).

C   Compliance

We found that there were 6 sessions with four different Control tutors who had used Tutor CoPilot in our current sample of sessions. While the 6 sessions make up an insignificant proportion of the total sample of 4,136 sessions in our current manuscript, this prompted us to investigate how the Control Tutors got access to Tutor CoPilot. We discovered that 6 Control Tutors were mistakenly assigned to Treatment within the tutoring provider. This does not change the balance between Treatment and Control, and for our results section, we will report all results based on our original assignment to the tutor.

D   Tutor CoPilot Training for Treatment Tutors

Before integrating Tutor CoPilot into live tutoring sessions, we provided treatment tutors with targeted training to ensure they understood its capabilities and how to use the tool effectively. Proper training was critical to avoid the common pitfall in education where new technologies are introduced without adequate instruction, leading to misinterpretation, misuse, and potentially negative impacts on students.[7] Without proper guidance, educators may develop incorrect expectations of the tool or use it in ways that hinder learning outcomes.

To prepare tutors, we provided the following training materials. We developed a slide deck illustrating various real-world tutoring scenarios that Tutor CoPilot supports and tutors went through this slide deck. Each scenario featured a student mistake, the lesson topic, a brief commentary on the nature of the mistake, and an expert strategy response generated by Tutor CoPilot. These examples are from actual tutoring interactions to make the training practical and relatable. Afterwards, treatment tutors were paired in a buddy system with another treatment tutor who would also get access to Tutor CoPilot. With their pair, they role-played in the "student" and "tutor" role. When a tutor played the "tutor" role, they had access to Tutor CoPilot and could see how Tutor CoPilot responded in real-time. The entire training process took approximately 2-3 weeks for all treatment tutors to complete.

---

[7]See, for example, the "Education and AI: Achieving equity and respecting the rights of students" event hosted by the Brookings Institution: https://www.brookings.edu/events/education-and-ai-achieving-equity-and-respecting-the-rights-of-students/.

E Study data

Tutors. For tutors, we received information on their gender (Male or Female), their pre-study quality rating (continuous between -1 and 1), and their pre-study tutoring experience counted as the number of months they've worked with the tutoring provider. The tutor's quality rating is determined by the tutoring provider based on human observations of the tutor's tutoring sessions and observation rubric scores. These scores are averaged to produce the quality rating used in our study. We performed tutor-level randomization and confirmed that these covariates are balanced between groups. We report the balance checks in Appendix B.

Students. For students, we received information on their gender (Male or Female), Race/Ethnicity (Hispanic, White, Black, Asian, Pacific Islander, American Indian or Alaska Native, Two or more races), whether they receive Free and Reduced lunch (Yes or No), whether they are in the Limited English Proficiency program (Yes or No) and their pre-study baseline and post-study NWEA MAP Math and Reading Scores. The NWEA MAP is a standardized test administered three times a year, tracking students' academic growth over time. For more information on NWEA MAP, please refer to https://www.nwea.org/map-growth/.

Sessions. We receive all the session-level data from our two-month study, which started at the end of March. We excluded sessions that were conducted with part-time tutors, as classified by the tutoring provider. The session-level data included the session's chat transcript, whiteboard activity, post-session survey responses from the student and tutor, the tutor's treatment assignment, and Tutor CoPilot use if applicable. The Tutor CoPilot use includes information on the session id, the request timestamp, conversation context preceding the click, the expert strategy, and corresponding suggestion generated by Tutor CoPilot. Our study includes 2,000+ uses of Tutor CoPilot.

F Exit Ticket Significance

Table 6 reports the regression results examining the relationship between students' exit ticket passing rate and their post-study test performance, while controlling for their pre-study baseline performance. We find that the exit ticket passing rate is statistically significant in predicting students' end-of-year test scores, suggesting that improvements in short-term learning outcomes measured by the exit ticket are associated with improvements in long-term academic performance.

Dependent	EOY Math Test Score
Exit Ticket Passing Rate	0.06***
	(0.01)
MOY Math Test Score	0.93***
	(0.01)
R^2:	0.864
n	959

Table 6: *Note.* $^+p < 0.1$; $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Regression results of a student's exit ticket passing rate predicting their post-study end-of-year (EOY) test performance, while controlling for the student's pre-study mid-of-year (MOY) test performance.

Panel A. Elementary Schools			
	Grade 3	Grade 4	Grade 5	Total
Total	676	1,828	357	2,861
Panel B. Middle Schools			
	Grade 6	Grade 7	Grade 8	Total
Total	1,275	0	0	1,275
Total				**4,136**

Table 7: Session-level statistics by grade. We had 8 elementary schools and 1 middle school in our study sample.

G   Descriptive Statistics

Study in numbers   Table 7 reports the session-level statistics. A majority of sessions are with elementary school students, particularly those in Grade 4. Table 8 reports student- and tutor-level statistics.

Tutor CoPilot usage   Figure 4 reports the percentage of treatment sessions that used Tutor CoPilot at least once in their session. On average, we find that about 29% of treatment sessions used Tutor CoPilot. Figure 5 reports the average number of uses per session, either (a) including or (b) excluding the sessions with zero uses. When including the zero-use sessions, tutors use Tutor CoPilot about 3 times per session. When excluding the zero-use sessions, tutors use Tutor CoPilot about 10 times per session. Overall, we find these high usage patterns suggest that tutors adopted the tool and integrated it well into their workflow. We did not send any reminders to tutors to use Tutor CoPilot.
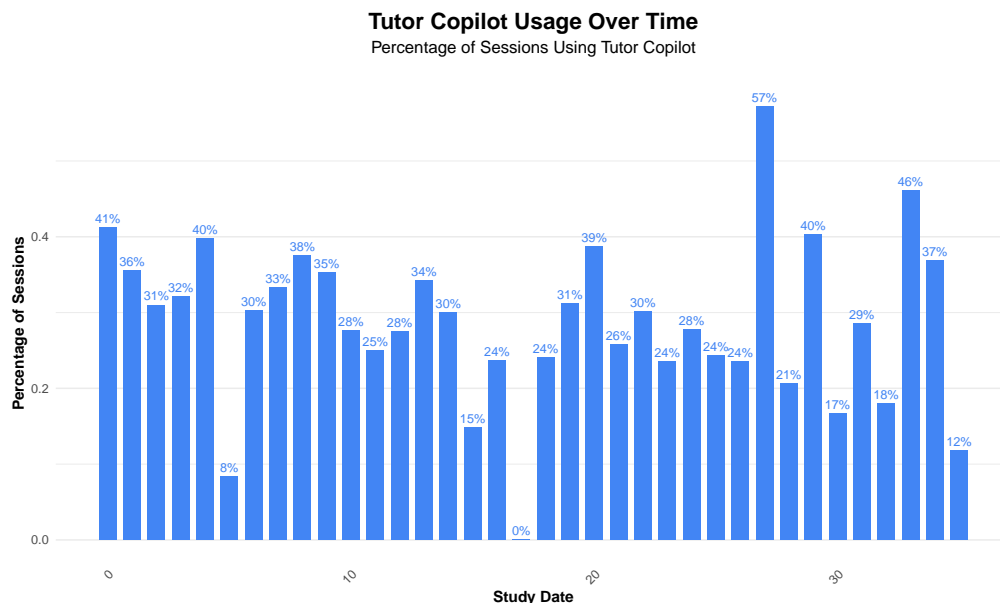

Figure 4: Percentage of treatment sessions that used Tutor CoPilot at least once in their session. About 29% of treatment sessions used Tutor CoPilot during our study.

		Mean
Student Characteristics		
Gender		
	Male	0.47
	Female	0.46
	Missing	0.07
Race/Ethnicity		
	Hispanic	0.805
	White	0.080
	Black	0.035
	Asian	0.003
	Pacific Islander	0.002
	American Indian or Alaska Native	0.000
	Two or more races	0.009
	Missing	0.066
Economically Disadvantaged		
	No	0.33
	Yes	0.67
In Limited English Proficiency Program		
	No	0.63
	Yes	0.31
	Missing	0.06
Mid-of-year Math MAP score		201.4
Tutor Characteristics		
Gender		
	Male	0.46
	Female	0.54
Quality Rating		0.41
Experience		21.5

Table 8: Student-level and tutor-level statistics.

Moments of Tutor CoPilot usage   We were interested in the moments of tutoring that tutors used Tutor CoPilot, e.g., did tutors use Tutor CoPilot at the start of the tutoring session or as the student started to solve problems? To study this, we developed a taxonomy of tutoring moments and trained binary classifiers to identify these tutoring moments. Appendix I details development process in more depth. Table 9 describes the taxonomy of tutoring moments, the frequency of these moments in our final data, and the test F1 score of the corresponding moment classifier. We find that tutors use Tutor CoPilot mostly during the moments of student learning: when the student is attempting the problem, or after they have attempted the problem and the tutor is giving them feedback.

Tutor adoption   Which tutors adopted the tool the most? In Table 10, we report the coefficients on tutor covariates predicting whether the tutor used Tutor CoPilot and number of uses in a given session. We also clustered the errors by tutor. We found that tutors who have more tutoring experience (measured by how long they've been with the tutoring provider) are less likely to use Tutor CoPilot, and female tutors use Tutor CoPilot one click more than male tutors.

Average Number of Tutor Copilot Uses Over Time
Included Zero–Use Sessions
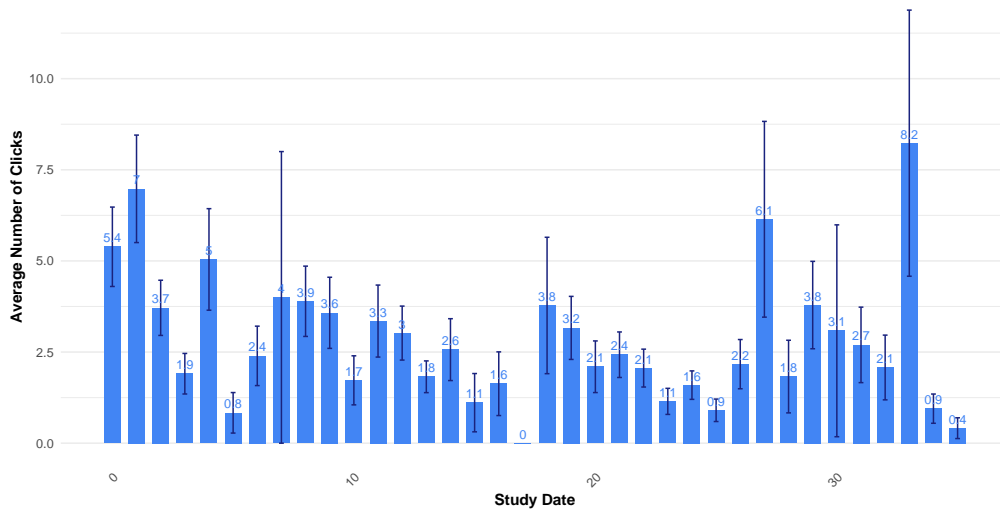
(a)

Average Number of Tutor Copilot Uses Over Time
Excluded Zero–Use Sessions
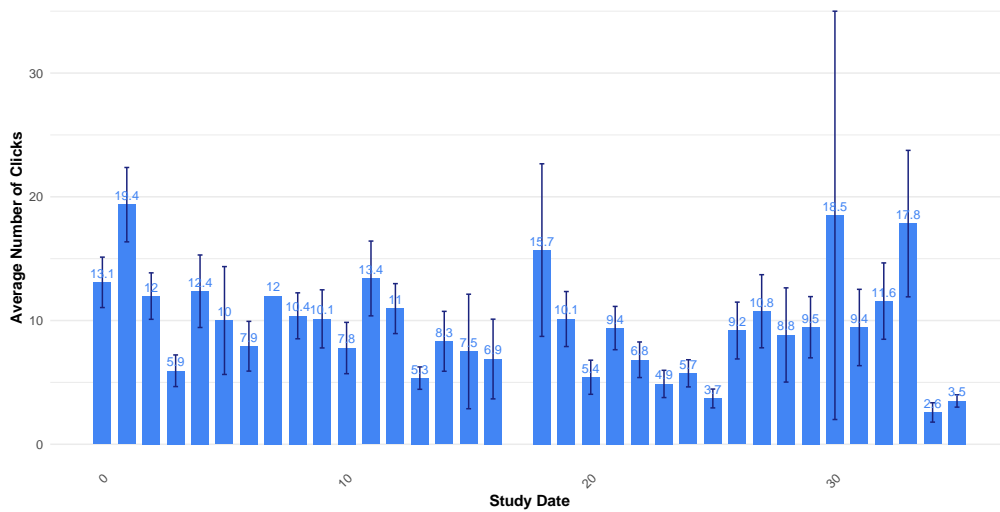
(b)

Figure 5: (a) reports the average number of Tutor CoPilot uses, including the sessions that had no usage, and (b) reports the average number of uses but excludes the sessions with no usage. When including the zero-use sessions, tutors use Tutor CoPilot about 3 times per session. When excluding the zero-use sessions, tutors use Tutor CoPilot about 10 times per session.

H   Treatment on the Treated Analysis

Setup   The ITT analysis instruments the tutor's treatment assignment to predict outcomes—but, the assignment alone may not capture the full picture. Tutors in some sessions may not use Tutor

Moment (Frequency)	Definition	Examples	F1
Start of session (0.9%)	The tutoring session is just starting. The student and tutor have not yet started a problem.	"Happy to work with you today!"	0.79
Start of problem (3.2%)	The tutor starts a new problem and/or gives instructions for the new problem.	"Go ahead and start showing your work for this question."	0.70
During problem attempt (49.5%)	The student is attempting the problem and/or the tutor has not yet given away the answer or explanation.	"Are you working on this problem?"	0.70
After problem attempt (42.1%)	The student has attempted the problem and the tutor is providing feedback. After a problem has been attempted, the tutor may want to start a new problem (category "start of problem").	"We know that we cannot subtract 1 - 3, so we will need to borrow from the whole number."	0.84
Start of exit ticket (1.1%)	The tutor starts an exit ticket for the student. Note that the exit ticket is a brief assessment near the end of the tutoring session and the tutor cannot help the student here, unlike for the normal problems.	"Now it is time for you to show what you have learned by completing the Exit Ticket."	0.83
During exit ticket attempt	The student is attempting the exit ticket.	"I can't help you with the exit ticket question."	0.0
After exit ticket attempt (4.6%)	The student has attempted the exit ticket and the tutor is providing feedback. Afterwards, the tutor may want to start a new exit ticket (category "start of exit ticket").	Congratulations. You've scored 100% in Exit Ticket questions."	0.90
End of session (1.9%)	The tutoring session is ending.	"We will continue in the next session."	0.98

Table 9: Taxonomy of tutoring moments, including their definitions, examples, and frequency over the labelled dataset. We train binary classifiers to identify these moments at scale and report their test F1 score as well. A majority of Tutor CoPilot usage concetrates during the "meat" of student learning: when the student is attempting the problem, or after they have attempted the problem and the tutor is giving them feedback. The classifier for "after exit ticket attempt" scored a low test F1 score, even after tuning the class-imbalance loss, thus we omit its frequency. Note that the frequencies do not sum to 1 because the classifiers are not mutually exclusive.

CoPilot, while in other sessions they may use it. To disentangle the effect of merely having access to the tool from the effect of actually using it, we extend our analysis to a treatment on the treated analysis (TOT) using a two-stage least squares (2SLS) regression. Intuitively, the 2SLS regression works as follows: In the first stage, we predict the likelihood of a tutor using Tutor CoPilot during a session based on their treatment assignment. This step isolates the variation in tool usage that can be attributed to being in the treatment group. In the second stage, we use this predicted likelihood to estimate the impact on the outcomes of interest. This method helps us understand the causal effect of using Tutor CoPilot, beyond just having access to it—or, in more technical terms, this method allows us to estimate the treatment-on-the-treated (ToT) effect, which specifically measures the impact of actually using Tutor CoPilot on the outcomes.

Concrete, we apply a similar model structure in the 2SLS framework to estimate the ToT effect. In the first stage, the treatment assignment, Treatment_j, is used to predict whether the tutor used Tutor CoPilot, Used_j. This predicted use is then used in the second stage to predict the outcomes:

Dependent Variable	Used	Number of Uses
Gender (Is Female)	0.01	1.36**
	(0.03)	(0.51)
Experience (Months)	-0.002+	-0.03+
	(0.001)	(0.02)
Quality Rating	0.12	1.69
	(0.08)	(1.41)
n	2014	2014

Note. $^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Estimates are controlled for the above variables and errors are clustered by tutor id.

Table 10: Predicting Tutor CoPilot use from tutor covariates. We report the coefficient estimates and the parentheses report the standard error.

$$\text{Used}_{ij} = \alpha + \beta_1 \text{Treatment}_j + \gamma X_i + \omega_k + \epsilon_{ij} \tag{2}$$

$$Y_{ij} = \alpha + \beta \text{Used}_j + \lambda X_i + \omega_k + \epsilon_{ij} \tag{3}$$

Findings We report the TOT findings in Table 11. We find a much larger, significant effect of treatment on our main outcome variables: Students are 14 p.p. more likely to pass their exit tickets when tutor use Tutor CoPilot (p<0.01), resulting in a 62% → **76%** passing rate. We also find that the exit ticket attempt rate increases by 6 p.p.

I Development and Classification with NLP classifiers

To analyze the use and impact of Tutor CoPilot, we want to identify the tutoring moments in which the tool was used and the pedagogical strategies used by tutors, as measured in their language. However, identifying these features requires overcoming several technical challenges. First, the chat transcripts, which span over 550,000 messages, are *unlabeled* for the moments and strategies. Therefore, we need a scalable and consistent method to label the data efficiently. Second, some categories, such as strategies that we consider high-quality, may be underrepresented in the data. Therefore, we need methods that can handle long-tailed distributions to handle rare categories. Finally, some tutoring contexts may involve actions that are not directly observable in the chat transcripts, such as students working on a whiteboard, leading to incomplete data. Therefore, the final method must also account for these contextual gaps. To address these challenges, we developed novel natural language processing (NLP) methods capable of efficiently labeling our entire dataset. Our approach involved three key steps:

1. **Taxonomy Development with Unsupervised Methods:** We began by creating a novel taxonomy of tutoring moments and pedagogical strategies. We constructed the taxonomies leveraging unsupervised methods (e.g., topic modeling) to efficiently organize the data and inform the development of the taxonomy. These taxonomies serve as the foundation for our analysis.

2. **Classifier Dataset Construction and Training for Handling Imbalanced, Imperfect Data:** Next, we label a dataset of 3,000 examples, annotated according to our taxonomies; we will refer to this

	Panel A. Session outcomes				
Dependent Variable:	**Participation Points**	**Participation Points (Standardized)**	**Exit Ticket Attempted**	**Exit Ticket Passed (Conditional)**	**Exit Ticket Passed (Unconditional)**
Treatment Effect	0.32	0.035	0.06^+	0.10^*	0.14^{**}
	(0.93)	(0.095)	(0.04)	(0.05)	(0.05)
Control Mean	14.07	0.016	0.84	0.73	0.62
	(0.20)	(0.02)	(0.01)	(0.01)	(0.01)
n	4136	4136	4136	3521	4136
	Panel B. Student survey outcomes				
Dependent Variable:	**My tutor cared understanding math over memorizing the solution.**	**My tutor cared about how well I do in math.**	**Even when math is hard, I know I can learn it.**	**Session Rating**	**Tutor Rating**
Treatment Effect	-0.01	0.09	0.06	-0.005	0.09
	(0.19)	(0.18)	(0.19)	(0.127)	(0.13)
Control Mean	4.19	4.31	4.24	4.77	4.74
	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)
n	1931	1931	1931	1948	1952

Note. $^+p < 0.1$; $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Estimates are from our primary model, which controls for baseline math scores, student demographics, and a fixed effect for strata (school $\times$ grade). Participation points are standardized within-sample and by grade. The survey items are on a 5-point scale where higher is better.

Table 11: TOT analysis on student session-level outcomes. The parentheses report the standard error.

    dataset as our *classifier dataset*. We train and evaluate the classifiers on this data, leveraging the imbalanced nature of categories to automatically reweigh our data and contextual information to mitigate the effects of incomplete data.

3. **Application of Classifiers on Downstream Datasets:** We run inference with the classifiers on our *downstream datasets*. We apply the moments classifier on our Tutor CoPilot usage data, which includes the conversation context leading up to the click of Tutor CoPilot. We apply the strategies classifier to both treatment and control chat sessions; we measure the causal impact of Tutor CoPilot on language by comparing the prevalence of strategies between groups.

The process of developing the taxonomy and training the classifiers is consistent across both moments and strategies; this process was agnostic to the treatment assignment condition. Note that the application of these classifiers differs based on the downstream datasets they are applied to. We will now elaborate on these steps in more detail.

I.1 Taxonomy Development

We initially developed a taxonomy for tutoring moments and strategies based on prior research and language patterns observed in tutors and students on the platform. To refine this taxonomy, we used the BERTopic package (Grootendorst, 2022), which allowed us to organize the diverse language used and clearly define the boundaries between categories. This process ensured that our taxonomy accurately captured the key distinctions needed for effective analysis. For both taxonomies, we used

`all-MiniLM-L6-v2` as our pretrained embedding model, and a count vectorizer on bigrams and trigrams as our vectorizer model; from qualitative experiences with tutoring language, we omitted unigrams because they tend not to cluster the text in pedagogically interesting ways.

Taxonomy of tutoring moments. We structured the taxonomy around the flow of tutoring sessions, beginning with typical phases: starting the session, working on practice problems, and completing exit tickets. Because we are particularly interested in moments of learning and tutor support—such as distinguishing moments when the tutor introduces the problems from moments when the tutor supports the student's attempt at the problem—we refined the moments into "start," "during student attempt," and "after student attempt" to capture these distinctions.

We further refined the definitions and boundaries of these moments by running BERTopic across the entire dataset. While this unsupervised approach highlighted useful patterns, it struggled with less common language, context-sensitive utterances, or unique ways that tutors engaged with students. Because we cared about capturing these patterns as well, we took a supervised learning approach to classify the tutoring moments more effectively but used the emergent topics to inform the taxonomy. Our final taxonomy of tutoring moments is shown in Table 9.

Taxonomy of strategies. We constructed the taxonomy of strategies first by using the strategies preferred and dispreferred by experts from prior work Wang et al. (2024a). The strategies preferred by experts were "explain a concept", "ask a question", "provide a hint", "provide a problem-solving strategy", "provide a worked example", "provide a minor correction", "provide a similar problem", "simplify the question", "affirm the correct answer" and "encourage the student". The strategies dis-preferred by experts were "give away the answer/explanation" or just asking the student to "recheck/retry" without any further guidance.

Again, we used topic modeling to refine the boundaries of these categories. For example, we realized that the "ask a question" category was too broad. Questions vary a lot in their pedagogical quality and usefulness. Questions can be used to merely check for understanding, like "Did you understand my explanation?" and questions can also be used to guide the student's thinking and be more process-oriented, such as "Which place value should be compared?" Based on these insights complemented by prior literature, we refined the "ask question" category to focus on questions that actively guided the student's thought process.

Not every message can be categorized into these strategies. Thus, we also had an N/A category, to capture things that should be ignored. This category included:

- Transition language: e.g., Let's do the next problem. or It's time to show your mastery in today's session.
- Checking in with the student: e.g., Are you there? or Let's focus on the session.
- Starting or ending the session: e.g., Let's start the session.
- Rushing the student: e.g., We are running out of time.
- Small talk: e.g., How was your day?
- Talk related to points: e.g., You receive an additional point for your efforts.
- Talk related to the platform: e.g., Let me increase my pace.
- Question instructions: e.g., In this question, you need to find the ordered pair represents a vertex of the parallelogram that is a reflection of Point S across the y-axis.
- Prompts: e.g., "Let me know if you need any help along the way." or "Give it a try."

I.2 Classifier Dataset Construction and Training on Imbalanced, Imperfect Data

To annotate our downstream dataset, we need to train classifiers on a subset of the data for the classifiers to reliably label for the moments and strategies. First, we subsampled a dataset of 3,000 examples that had a balanced representation from both treatment and control chat transcripts. Each example consisted of a pair of the context (10 prior messages) and the tutor's message following this context.

To efficiently label the classifier dataset, we adopted a Human+LLM annotation approach. This method follows recent trends in combining human expertise with AI to scale annotation efforts (Kim et al., 2024, Wang et al., 2024b). We prompted an LLM to first annotate the dataset using our taxonomies, after which human annotators (two co-authors) reviewed and corrected the labels. Unlike other studies on Human+LLM annotation approaches that start with an existing codebook, our codebook combined unsupervised methods to inform the taxonomy developed as previously described.

We split the annotated dataset for both moments and strategies into training, validation, and test sets in a 6:1:3 ratio. We cast this setting as a multiple binary classification task; this provides multiple advantages, such as this doesn't assume mutual exclusiveness among classes, which aligns well with real-world data where few classes might be similar with each other, and each class is considered independent with its own predictor. This is a nice property since real-world data often has more than one semantic label.

We finetuned a RoBERTa large model and introduced new tokens to separate the context and target text to be classified: `[CONTEXT_TOKEN] {context} [TARGET_TOKEN] {utterance}`. Given the long-tailed nature of some categories, we automatically reweighed the data based on class distributions and optimized the classifier using a sigmoid cross-entropy class-balanced loss (Cui et al., 2019). The class-balanced loss introduces a weighting factor that is inversely proportional to the effective number of samples needed to train a good classifier. We ran a hyper-parameter sweep over the the loss' hyper-parameters and learning rate on the validation set, selecting the hyper-parameters that yielded the lowest validation loss. The performance of the classifiers is reported using the F1 score on the test set, shown in the taxonomy tables in the main paper.

We set a priori thresholds on the test F1 scores of 0.60+ for categories to include in our downstream analysis. Categories that did not meet this threshold were excluded from further analysis. This led us to dropping "during exit ticket attempt" from the moments analysis, and "explanation of concept" and "provide hint" from the strategies analysis.

I.3 Application of Classifiers on Downstream Datasets

After training and validating our classifiers, we applied them to our downstream datasets to answer our research questions.

Moments. To understand the tutoring moments in which treatment tutors used Tutor CoPilot, we run inference with our moment classifiers on the Tutor CoPilot usage data. We use this dataset to identify the conversation context leading up to the tutor's click on the tool and the message that followed. We formatted this data the same as the training format. After running inference on the dataset, we report the frequency of these identified moments within the Tutor CoPilot usage data. By doing so, we can tied these frequency findings back to our research question on the type of tutoring moments tutors used Tutor CoPilot.

Strategies. To measure the the causal language impact of Tutor CoPilot, we run inference with our strategy classifiers on all chat transcripts. We then separate the transcripts into treatment and control groups based on the tutor's treatment assignment, and identify which strategies are more prevalent in one group over the other. To quantify these differences, we adapted the Fightin' Words method (Monroe et al., 2008) which calculated the z-score log odds to compare frequency of strategies while adjusting for a prior distribution. This method quantifies the causal differences in strategies between treatment and control tutors. We report the z-scored log odds for each strategy in Figure 3.

J   Structured Interview Protocol with Tutors

To understand how tutors perceive the Tutor CoPilot tool, we conducted an interview following a structured interview protocol a week after the study had concluded. The interview was done virtually. The main author of this work was the discussion lead, and the discussion included approximately 20 tutors, alongside members of the tutoring operations team and an engineering manager. The structured format was chosen to facilitate more focused and productive discussions, as it can be challenging for tutors to provide feedback in an unstructured group setting.

The interviews were conducted in a 1-hour session, following a specific protocol:

1. **Session Timing**: The session was scheduled to last one hour. We allowed the first five minutes for everyone to join, with the discussion beginning promptly at :05.

2. **Recording Consent**: At :05, the discussion lead asked participants if they were comfortable with the session being recorded on Zoom. Once consent was obtained, the session was recorded to ensure accurate capture of the feedback provided.

3. **Session Introduction**: The discussion lead communicated that the purpose of the session was to gather their feedback on Tutor CoPilot. The lead emphasized that the session was not an evaluation of the tutors, but rather an opportunity for the tutors to share their experiences and thoughts on Tutor CoPilot. This would inform how future iterations of Tutor CoPilot could be improved.

4. **Structured Questions**: Participants were provided with a list of questions in a shared Google document. The questions were:
   - Have you used Tutor CoPilot? If so, how?
   - Can you provide an example of how you've used it?
   - How has your use of Tutor CoPilot changed over time?
   - What aspects of Tutor CoPilot do you like? What do you not like?
   - What would you like Tutor CoPilot to do for you? Are there specific times during tutoring sessions when you would like additional support from the CoPilot?
   - Do you have any other thoughts or feedback on Tutor CoPilot?

5. **Reflection Period**: Participants were given seven minutes, ending at :20, to reflect on the questions and note down their responses in the shared document.

6. **Group Discussion**: Following the reflection period, we went through the responses together, allowing for a discussion of the feedback shared. This part of the session was intended to foster a collaborative exchange of ideas and experiences among the tutors.

7. **Conclusion**: The session concluded with a thank you to all participants for their input. They were also encouraged to send any additional thoughts or questions to a provided email address.

K Student-level analysis

In this analysis, we measure the effect of treatment exposure on student end-of-year test scores. We instrument treatment exposure as the proportion of the number of tutoring sessions with a treatment tutor over the total number of tutoring sessions attended by the student. To conduct a student-level analysis, we removed students from our sample who did not participate in any tutoring sessions during the study period. Thus, our analytical sample only includes students with exposure to tutoring, and thus potentially to Tutor CoPilot.

Treatment exposure   Figure 6 reports the histogram on the treatment exposure as a proportion. We find that few students had high exposure to Tutor CoPilot, and overall, there was limited variation in treatment exposure across the student population, with most students showing near-zero exposure. We therefore expected that the impact on end-of-the-year outcomes would not be significant.
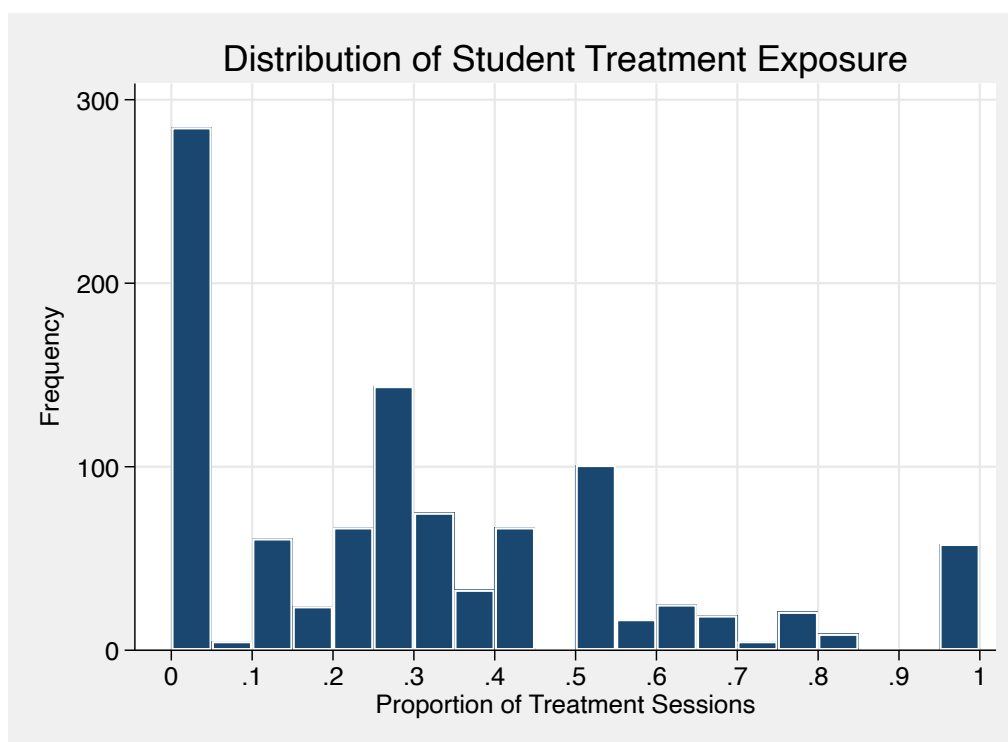

Figure 6: Histogram of student treatment exposure.

Regression findings   Table 9 reports the estimates of the treatment exposure on the student's test scores in math and reading. We find no significant effect on the test scores, though with math scores trending negative and reading scores trending positive. The lack of significance may be attributed to the limited variation in treatment exposure among students and insufficient exposure to the treatment, which likely constrained the potential for detecting an effect.

L Tutor-level analysis

The tutor-level analysis is conducted on their end-of-study survey. We first check for balance on the response rate of the survey, then the estimates of treatment on the tutor-level outcomes.

Dependent Variable	With Imputed Baseline		With Non-Imputed Baseline	
	Math MAP	Reading MAP	Math MAP	Reading MAP
Treatment Exposure	-0.35	1.44	-0.59	1.56
	(0.88)	(1.45)	(0.86)	(1.42)
n	1,001	1,002	959	959

Note. $^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Estimates are from our primary model, which controls for baseline math scores, student demographics, and a fixed effect for strata (school $\times$ grade).

Table 12: Student-level regression. The parentheses report the standard error. We report the coefficient estimates both with the imputed test baseline and the non-imputed test baseline for each student.

Balance check on response rate   Table 13 reports the tutor's response rate to the end-of-study survey. We saw no difference in response rate across assignment groups.

	Coefficient	*p-value*
Treatment	-0.01	NS
	(0.02)	
Gender (Is Female)	0.02	NS
	(0.02)	
Experience (Months)	0.001	NS
	(0.001)	
Quality Rating	0.09	NS
	(0.06)	

Note. NS = Not Significant. Experience reports the number of months the tutor has been with the tutoring provider. Quality Rating reports the tutor's tutoring quality rating determined prior to the study by the tutoring provider based on session observations & quality rubric scores.

Table 13: Balance check on tutor response rate for the end-of-study survey.

Post-study survey results   Table 14 reports the treatment estimates on the tutor-level outcomes from the post-study survey. We find no significant changes between the treatment and control groups.

M   Study Costs

The total API cost for 429 treatment tutors over the 2-month study was $1,419.66, resulting in an estimated annual cost of $20 per tutor.

Dependent Variable	Do you agree or disagree with this statement: "My students learn more by making mistakes."?	How confident are you at recognizing the kind of mathematical mistakes students are making?	How effective are you at helping students fix their mistakes?	How much more or less effective are you at helping students fix their mistakes now than you were three months ago?
Range	1 = Disagree	1 = Not at all confident	1 = Not at all effective	1 = A lot less effective
	2 = Somewhat disagree	2 = Slightly confident	2 = Slightly effective	2 = Slightly less effective
	3 = Somewhat agree	3 = Confident	3 = Effective	3 = No change
	4 = Agree	4 = Very Confident	4 = Very effective	4 = Slightly more effective
				5 = A lot more effective
Treatment	-0.02	0.05	0.03	0.005
Effect	(0.06)	(0.06)	(0.03)	(0.05)
Control	3.48	3.37	3.81	4.60
Mean	0.04	0.04	0.02	0.04
n	766	766	766	766

Note. $^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Estimates are from our primary model, which controls for tutor covariates.

Table 14: Tutor-level outcomes. The parentheses report the standard error.