AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms

LearnLM Team, Google & Eedi

One-to-one tutoring is widely considered the gold standard for personalized education, yet it remains prohibitively expensive to scale. To evaluate whether generative AI might help expand access to this resource, we conducted an exploratory randomized controlled trial (RCT) with N = 165 students across five UK secondary schools. We integrated LearnLM—a generative AI model fine-tuned for pedagogy—into chat-based tutoring sessions on the Eedi mathematics platform. In the RCT, expert tutors directly supervised LearnLM, with the remit to revise each message it drafted until they would be satisfied sending it themselves. LearnLM proved to be a reliable source of pedagogical instruction, with supervising tutors approving 76.4% of its drafted messages making zero or minimal edits (i.e., changing only one or two characters). This translated into effective tutoring support: students guided by LearnLM performed at least as well as students chatting with human tutors on each learning outcome we measured. In fact, students who received support from LearnLM were 5.5 percentage points more likely to solve novel problems on subsequent topics (with a success rate of 66.2%) than those who received tutoring from human tutors alone (rate of 60.7%). In interviews, tutors highlighted LearnLM's strength at drafting Socratic questions that encouraged deeper reflection from students, with multiple tutors even reporting that they learned new pedagogical practices from the model. Overall, our results suggest that pedagogically fine-tuned AI tutoring systems may play a promising role in delivering effective, individualized learning support at scale.

Keywords: learning, efficacy, safety, artificial intelligence, tutoring, randomized controlled trial

1. Introduction

One-to-one tutoring is the gold standard for supporting students' learning and education. Decades of research demonstrate that individualized tutoring results in substantial gains in learning [1–3]. Unfortunately, the high cost of one-to-one tutoring and relative scarcity of educators makes this support inaccessible for most students and classrooms. The tension between tutoring's effectiveness and inaccessibility presents an enduring challenge for education systems: can educators deliver individualized support in a way that is both highly effective and broadly scalable?

A growing number of researchers and practitioners now look to generative AI ("genAI") as a potential solution to this challenge [4–7]. Indeed, a wave of new tutoring systems incorporate genAI for direct interactivity with students [8]. Yet rigorous, in-classroom research on the learning efficacy of genAI remains scarce [9]. The evidence that does exist is mixed: while some studies suggest genAI can offer effective instruction [10–13], others find that deploying genAI tutoring systems without appropriate pedagogical safeguards can actively harm learning [14, 15].

Here we report the results of an exploratory randomized controlled trial (RCT) with N = 165 students, designed specifically to evaluate if an AI tutor can safely and effectively support students in UK secondary school classrooms. Our study took place on the Eedi educational platform, an evidence-based learning ecosystem that provides students with both curriculum-aligned mathematics activities and one-to-one support from remote human tutors via online chat conversations. In our experiment, we tested whether LearnLM—a genAI model fine-tuned for pedagogical applications [16–18]—could help scale this assistance.[1] To ensure a high standard

[1] Since the launch of this RCT, Google DeepMind has infused the pedagogical capabilities of LearnLM into the Gemini model family (see goo.gle/learnlm for details). Teams seeking to replicate these findings or build similar experiences should now use Gemini 2.5 Pro.

Corresponding author(s): learnlm-tech-report@google.com

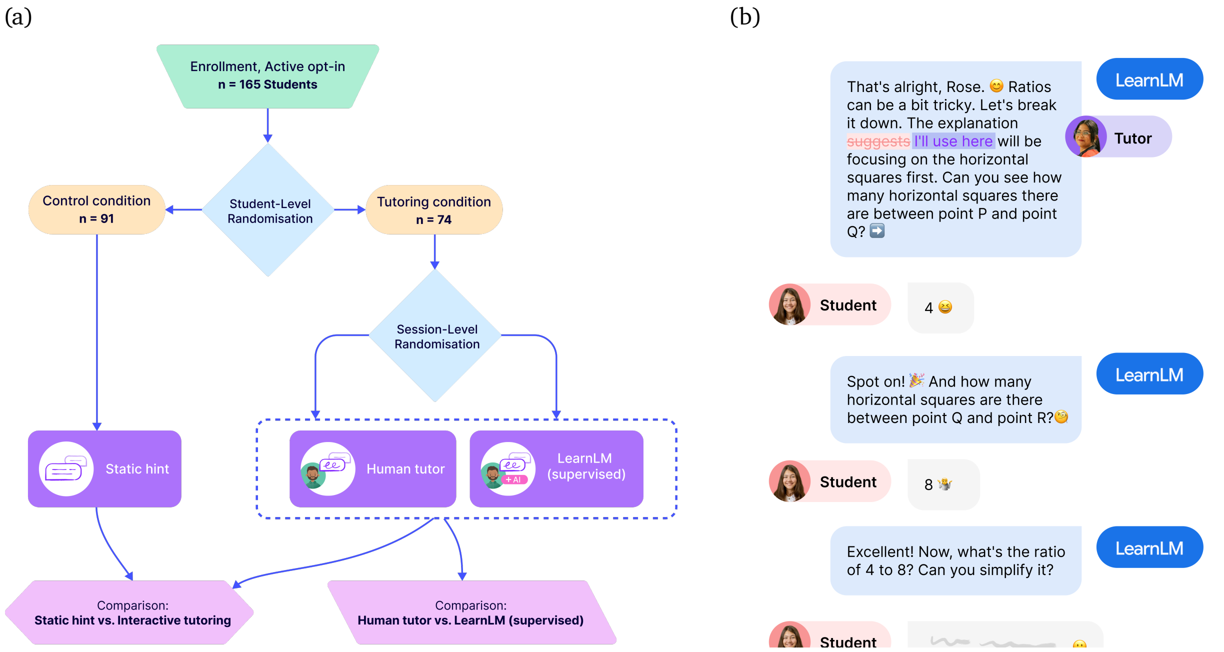

Figure 1 | We designed this exploratory RCT to evaluate the safety, pedagogy, and efficacy of LearnLM. (a) The RCT randomly assigned each of $N = 165$ students to receive either static hints or interactive tutoring. Students in the tutoring condition experienced a further level of randomization. When they started a tutoring session, the platform randomly assigned them to either a session with a human tutor or a session with LearnLM (supervised by a human tutor). This design allows us to compare static, pre-written support against interactive tutoring, as well as human tutoring against (supervised) tutoring from LearnLM. (b) In sessions with LearnLM, a supervising tutor reviewed each message that LearnLM drafted. They could either edit the message, completely re-write it, or approve it without any changes. The Eedi platform then sent the message to the student.

of safety and pedagogy for all students in our trial, $N = 17$ expert human tutors directly supervised LearnLM, assuming ultimate responsibility for every interaction it had with students. In particular, the tutors appraised each message that LearnLM generated, retaining full control to approve, edit, or replace it before it reached the student.

LearnLM proved to be a trustworthy source of pedagogical instruction, with the supervising tutors approving over 76% of its messages without changes or with only minimal edits (changing one or two characters; e.g., deleting an emoji). In fact, across all of the learning outcomes we measured, supervised support from LearnLM proved at least as effective as guidance from a human tutor. Most surprisingly, students tutored by LearnLM exhibited measurably better knowledge transfer than those receiving support from human tutors alone. On average, supervised support from LearnLM improved the probability of a student solving a novel problem correctly by 5.5 percentage points over guidance from a human tutor.

To better understand this broad effectiveness, we surveyed and interviewed the supervising tutors for their perspectives on LearnLM. They reported that LearnLM consistently generated high-quality, Socratic dialogue, providing a strong foundation for academic interactions with students. The supervising tutors' interventions tended to focus on moderating the dialogue's pacing and providing the social and emotional nuance required to maintain student engagement.

Overall, our exploratory RCT identifies several avenues for new research on AI and education, while also suggesting a potential role for genAI tutors in delivering effective, individualized learning support at scale.

2. An Exploratory Classroom Trial

Our RCT aimed to evaluate LearnLM in a rigorous, real-world, in-classroom testbed. Hundreds of secondary schools in the UK integrate the Eedi learning platform directly into their mathematics instruction. The platform provides students with curriculum-aligned study units and a spectrum of personalized support, including two forms of assistance central to this RCT: carefully designed hints for common misconceptions in each study unit, and one-to-one guidance from trained, expert tutors via online chat interactions. Students who receive this standard support on the Eedi platform experience the equivalent of two additional months of academic progress, with the impact doubling for highly engaged students [19]. We recruited $N = 165$ students in Year 9 and 10 (ages 13–15) across five of these schools for the RCT (see Appendix A). Each student and each tutor provided informed consent to participate in the trial. The trial ran from May through June 2025.

The trial leveraged these two forms of Eedi support—hints and chat-based tutoring ("hybrid tutoring" [20])—as baselines to assess the pedagogical efficacy of LearnLM (see Figure B.1 in Appendix B). During the trial period, we randomly assigned each student either to receive *static* pedagogical support (pre-written hints) or to enter an *interactive* one-to-one tutoring session (Figure 1; see also Appendix B). Students in the tutoring condition experienced a further level of randomization: when a student entered a tutoring session, we randomly connected them either with an expert human tutor or with LearnLM (supervised by a human tutor). We prompted LearnLM to adopt a Socratic approach aimed at guiding the student to identify their own mistake, and provided the model access to the full question text, the student's incorrect answer, and explanations for both the student and a teacher about the misconception underlying the incorrect answer, among other information (see Appendix D.1).

Our approach allowed us to pose a set of four research questions:

RQ1: Was LearnLM a reliable and pedagogically sound source of instruction?
RQ2: Was interactive tutoring (whether delivered by a human tutor alone or in a supervised session with LearnLM) more effective for student learning than static pedagogical support?
RQ3: For students receiving interactive tutoring, was support from a supervised session with LearnLM more effective than support from a human tutor working alone?
RQ4: What can we learn from tutor and student experiences of interacting with LearnLM?

To answer these questions, we adopted a Bayesian framework and directly estimated the magnitude and credibility of our treatment effects. Unlike standard frequentist approaches, this method allows us to calculate the probability that one intervention outperforms another by a specific magnitude, providing a more practical foundation for making decisions about real-world deployment. For all analyses, we assigned identical, weakly informative priors to each intervention. We then used the resulting posterior distributions to calculate the exact probability that outcomes in one group exceeded those in another, providing a more precise signal than a simple comparison of the intervals (cf. [21–23]). For complete experimental details, see *Methods* and Appendices A–D.

3. Results

We first verified the basic safety and quality of LearnLM's tutoring (**RQ1**) by auditing the full corpus of 3,617 messages that it drafted, as well as the supervising tutors' decisions to approve, edit, or rewrite those messages. LearnLM proved a trustworthy source of instruction. The tutors who supervised and reviewed its messages accepted 74.4% without any edits. As judged by edit distance [24, 25], many of the $k = 926$ instances where tutors edited or rewrote a suggestion reflected minor or targeted adjustments (see Table E.1 for examples). The two most frequent edit distances, accounting for 5.5% and 2.4% of re-writes, were just a single character and two characters, respectively; these virtually always reflected a tutor deleting or changing an emoji. The median intervention altered 59 characters, or just a few words. Still, after the RCT finished, we asked the supervising tutors to systematically review the corpus of edits and re-writes. This review revealed zero instances of harmful or risky content and only five factual errors, or 0.1% of the total 3,617 messages that LearnLM drafted (see Table E.2 in Appendix E). Overall, a close audit confirmed that LearnLM provided safe and reliable guidance during the trial.

Next, we evaluated effects on student learning (**RQ2**, **RQ3**), comparing students' performance after receiving one of the standard interventions on the Eedi platform or interacting with LearnLM. As described in *Methods*, students worked through a series of short study units, each consisting of several multiple-choice questions

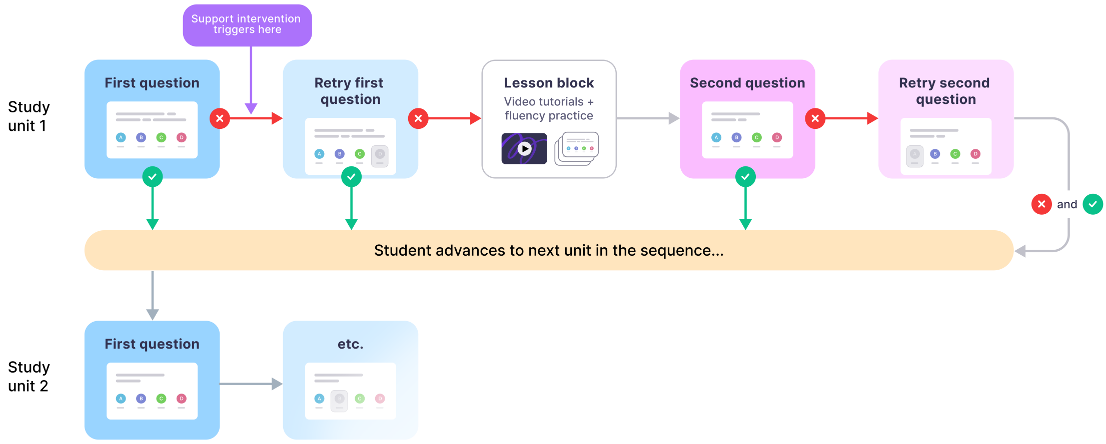

Figure 2 | Student progression through the study unit. If a student makes a mistake on the first question in a study unit, they receive a support intervention. We analyze whether the intervention helps the student identify and remediate their mistake, resolve the misconception underlying their incorrect choice, and transfer the knowledge from the intervention to the next study unit. See *Methods* and Appendix C for more information on the Eedi platform.

designed to assess a specific mathematical topic (Figure 2). Whenever a student answered the first question in a unit incorrectly, the platform triggered a support intervention. Depending on their assigned condition, students either received a static, pre-written hint specific to their mistake on that question, or an interactive (chat-based) session with a tutor. Immediately following the intervention, the platform presented the student with the exact same question and prompted them to try answering it again.

Echoing prior research [3], interactive support with a human tutor proved far more effective for this kind of immediate course-correction. Students who joined a real-time session with a human tutor were substantially more likely to correct their mistakes than were those who received a static, pre-written hint (see Figure 3, left). In particular, 91.2% of students who received interactive support from a human tutor solved the problem correctly on their second attempt (with a 95% credible interval of [88.5%, 93.6%]), compared to only 65.4% [63.8%, 66.9%] of students who received a static hint. Supervised instruction from LearnLM proved just as effective at helping students correct their mistakes. Students receiving guidance from LearnLM answered their second attempt correctly 93.0% [90.4%, 95.3%] of the time. (For context, simply eliminating the previous mistake and guessing from the remaining options would yield an expected success rate of 33.3%.)

If a student still answered the question incorrectly on their second attempt, the platform provided them with several additional opportunities to correct their underlying misconception. Specifically, it offered them two attempts at a new question on the exact same mathematical topic. We thus examined whether tutoring helped students eventually resolve their misunderstanding—that is, whether they answered *any* of the post-intervention questions correctly. On this broader measure, interactive tutoring once again proved superior to static hints (see Figure 3, center). When working with a human tutor, 94.9% [92.6%, 96.8%] of students resolved their misconception, relative to only 86.8% [85.7%, 88.0%] of students receiving pre-written hints. No meaningful difference emerged between students working with LearnLM and those working with human tutors. Students tutored by LearnLM resolved misconceptions 95.4% [93.1%, 97.1%] of the time. For this kind of near-term correction, both interactive methods appear equally effective.

Of course, the critical question is whether these guided successes (the opportunity to remediate mistakes and resolve misconceptions) reflect durable learning (the ability to solve a new problem without any assistance). Within the scope of this RCT, the best test for durable effects of tutoring is how students performed when progressing to a new topic. The Eedi platform organizes study units into sequences of five, where each unit builds directly upon the last. Our subsequent analysis therefore analyzed a student's likelihood of correctly answering the initial question in the very next unit in their current sequence.

Here, a clear advantage for LearnLM's tutoring emerged (see Figure 3, right). Students tutored by LearnLM

Figure 3 | Tutoring interventions improve student learning outcomes. (left, center) For immediate learning outcomes, sessions with human tutors and expert-supervised sessions with LearnLM promote similar growth for students. Students who receive interactive tutoring from either source substantially outperform students who receive pre-written, static hints. (right) In contrast, students tutored by LearnLM demonstrate greater knowledge transfer to new topics than those supported either by static hints or by human tutors alone. Error bars indicate 95% credible intervals. Dashed lines represent the chance of success when guessing randomly (33.3%, 66.7%, and 25%, respectively).

on a study unit proved substantially more likely to answer the first question in the following unit correctly (66.2% [61.1%, 71.2%]) than students who had received help from an unassisted human tutor (60.7% [55.8%, 65.4%]). In particular, a supervised session with LearnLM increased the likelihood of learning transfer to a distinct topic by an additional 5.5 percentage points [–1.4%, +12.4%] relative to human tutoring. Both tutee groups, in turn, outperformed students who had received only a static hint (56.2% [54.2%, 58.2%]). Altogether, we attribute a high credibility (93.6%) to LearnLM offering better support for knowledge transfer than human tutors alone, and near certainty (>99.9%) to its advantage over static, pre-written hints. The AI-supported interventions fostered a more durable and transferable understanding—an advantage revealed only when students faced a fresh challenge. (See Appendix F for our full analysis of learning outcome data.)

Throughout the RCT, we sought a richer, more nuanced understanding of the experience of interactions with LearnLM (**RQ4**) by conducting in-depth, semi-structured interviews with a random subset of $N = 5$ supervising tutors (see Table 1). In addition, we invited all students and supervising tutors to share their thoughts in brief surveys. We gathered $N = 27$ student responses from a post-trial survey, and $N = 17$ tutor responses on both pre- and post-trial surveys. These firsthand perspectives help contextualize LearnLM's effectiveness and the specific role that human expertise played in its tutoring successes.

Over the course of the trial, supervising tutors came to view LearnLM as a source of high-quality, expert-level pedagogical insights. The most prominent theme from our interviews, raised independently by all five interviewed tutors, was LearnLM's consistent use of Socratic dialogue. Tutors reported that its suggestions prompted a more inquisitive, student-led interaction. One tutor highlighted its ability to ask "really good questions that I hadn't necessarily thought of [...] in a good way, a nice way" (T3). As another reported, "[LearnLM] definitely explained certain topics in a better way than I probably could have" (T5). This praise aligned with tutors' actions during the trial: as established earlier, the tutors approved the vast majority of the

Tutor ID	Gender	Years of teaching experience
T1	F	6–10 years
T2	F	More than 10 years
T3	F	More than 10 years
T4	F	More than 10 years
T5	F	6–10 years

Table 1 | We conducted semi-structured interviews with a subset of five supervising tutors to seek a deeper, nuanced understanding of LearnLM's behavior and the general experience of participating in the RCT. Table A.1 in Appendix A contains comparable details for the full sample of supervising tutors.

messages drafted by LearnLM without any edits or changes.

In interviews, three tutors noted that supervising this high standard of instruction prompted an unexpected outcome on their part: professional growth and development. For instance, one tutor contrasted LearnLM's Socratic strategy with their prior approach, noting that the drafted messages prompted "questions more like 'Okay, what made you think that was the answer?' [...] whereas before [...] my main goal was to identify their misconception myself" (T1). Another explained, "I remember thinking, 'Oh, I hadn't thought of explaining it that way before.' Just like when you watch another teacher" (T2). Over the course of the trial, LearnLM's standard of instruction made a considerable impression on the tutors who supervised it.

One-to-one tutoring requires sustained, substantive effort to process the scenario at hand and craft effective pedagogical guidance. LearnLM's ability to consistently generate high-quality pedagogical responses thus made the entire tutoring process more fluid and efficient. Our post-trial surveys corroborated this; when asked about LearnLM's most useful feature, 82.4% of tutors chose "supporting multiple students at the same time." This new, effective process quickly set a new standard for the supervising tutors. In fact, every tutor that we interviewed independently raised this increased capacity as a key strength. As one tutor explained in their interview, "I got to the point of being disappointed when I didn't get [a session] with the AI suggestions" (T2). These positive experiences translated into a broad increase in comfort with AI across the cohort. Tutors' self-reported comfort with using AI tools rose from an average of 3.4 [2.9, 4.0] out of 5 in the pre-trial survey to 3.9 [3.3, 4.4] in the post-trial survey (posterior probability of increase: 90.0%).

Building LearnLM's pedagogical insights into effective tutoring conversations, however, required the supervising tutors to incorporate social and emotional nuance from their understanding of the students. Our retrospective analysis of the 25.6% of cases where tutors edited or re-wrote LearnLM's messages identified two primary motivations for these interventions: moderating the pedagogical pacing of the conversation and providing social-emotional nuance to LearnLM drafts. The most frequent intervention was adjusting the conversation's pacing to prevent exasperating students, accounting for 44.3% of all edits. Our tutors echoed this specific challenge in five of our five interviews. As one tutor explained in their interview, "quite often the students just got frustrated, and then they lost complete interest in the question, so it was a case of overriding it" (T2). Tutors often found it necessary to step in when LearnLM's Socratic questions, while pedagogically sound, persisted longer than a student's patience. One tutor described a common scenario where "[LearnLM] will go, 'Okay, you've got the answer. Let's dig a little deeper about why you've got that answer.' And the child is just like, 'No, I've got it. I know what I'm doing. Can I go now?'" (T1).

Providing social and emotional context to LearnLM's drafts emerged as a second prominent motivation for supervisors' interventions. In total, 19.5% of tutors' edits adjusted the persona or tone conveyed by the drafted messages. Tutors consistently added personal touches that recognized the student as an individual. For example, one tutor noted the importance of acknowledging a student they had helped before, a nuance LearnLM could not replicate, given that its prompt did not provide any information on past tutoring sessions: "...if you'd already helped that student twice before, [LearnLM] didn't quite have the capability to go like, 'Oh Sarah, it's you again. Hi!' And I like to have that kind of rapport" (T3). Tutors also calibrated the tone of the messages to ensure they were appropriate for student communication styles. One tutor remarked that LearnLM's predilection for emojis "comes across as a bit fake, and [...] the students pick up on that" (T1). Overall, the human tutors grounded LearnLM's suggestions with social and emotional nuance, translating its pedagogical insights into effective educational interactions.

Finally, student feedback indicated broad satisfaction with their tutoring interactions. In post-trial surveys, students who received interactive tutoring rated the helpfulness of the support they received an average of 3.9 [3.1, 4.7] out of 5, relative to 3.6 [2.9, 4.2] for students who received static hints (posterior probability of an advantage for tutoring: 74.9%). Ultimately, interactive tutoring delivered not just strong learning outcomes, but an enjoyable experience for the learners themselves.

4. Discussion

When deployed responsibly, can generative AI safely and effectively support students in real-world learning environments? Our exploratory trial investigated whether LearnLM—a genAI model fine-tuned for pedagogical applications—could help provide in-classroom guidance across five UK secondary schools. Students in these schools use Eedi, an online mathematics platform that effectively improves learning outcomes [19], for their regular instruction. We incorporated LearnLM into the platform so that it drafted messages to send to students in chat-based tutoring sessions. Of course, genAI tools carry well-known risks, including their capacity to fabricate information [26, 27] and erode critical thinking [28, 29]. Given the heightened ethical weight of these risks in educational settings, we assigned a group of expert (human) tutors to directly supervise LearnLM, assuming ultimate responsibility for each of its interactions with students. The tutors applied a simple, rigorous standard: they revised each of LearnLM's drafts until they were satisfied sending the message as their own.

The supervising tutors found LearnLM to be a reliable source of pedagogical instruction, approving the vast majority of its drafted messages without any edits. A systematic review of the drafted messages revealed zero instances of harmful content and only five factual errors out of 3,617 messages drafted by LearnLM total. For students, this translated into effective support for learning: tutoring from LearnLM helped students identify their mistakes and correct their misconceptions just as well as instruction from human tutors alone. Unexpectedly, students tutored by LearnLM demonstrated greater knowledge transfer to subsequent topics than did students who received guidance from human tutors.

Tutors consistently praised LearnLM's use of Socratic dialogue, but also noted that the model's relatively inflexible adherence to pedagogical principles threatened to exasperate some students. The best human tutors, in contrast, draw on experience, empathy, and judgment to decide when to push students and when to moderate their approach. This is a constant judgment call for tutors: weighing the long-term benefits of productive struggle against the immediate risks of frustrating a student and causing them to disengage completely. This delicate calibration remains a fundamental challenge for current AI systems [30–33].

Beyond safety and pedagogy, expanding access to one-to-one tutoring will require improving its cost and scalability. In our interviews, the supervising tutors consistently reported that LearnLM made their work feel more fluid and efficient. Our own anecdotal observations during the trial supported these reports: tutors appeared comfortable managing higher workloads during their supervised sessions. Unfortunately, the design of this RCT—with tutors fluidly switching from supervision to direct interaction during the same classroom periods—precludes a rigorous measurement of throughput or efficiency for each condition. After the trial, we simulated additional sessions as an informal test of scalability (see Appendix G). The results of this informal test corroborate the improved efficiency of the supervised sessions, with tutors sustaining a higher volume of simultaneous conversations when supported by LearnLM. Altogether, these signals support a possible role for genAI tutoring in helping educators to deliver individualized instruction at scale.

Overall, the design of this exploratory RCT allowed us to rapidly validate LearnLM's safety and gather initial signals of its efficacy. We measured these outcomes using students' standard, daily activities on the Eedi platform. This approach provided us with learning signals immediately, eliminating the need to develop and administer new trial-specific assessments, or to wait for the next round of standardized exams. In addition, by randomly assigning the source of support for each individual tutoring session, we could measure the alternating impact of LearnLM and human tutoring on the same students. This approach disentangled tutoring effectiveness from pre-existing student differences, permitting us to detect meaningful indications of efficacy working with just five schools.

On the other hand, this design offers only a partial glimpse at the broader trajectory of learning. Randomizing the source of tutoring session-by-session allowed our RCT to efficiently investigate immediate learning outcomes, but also prevented it from isolating the cumulative impact of working with LearnLM over time. Measuring substantive, longer-term effects on learning will require a different approach. In addition, the finding from our interviews that tutors learned from supervising LearnLM indicates another methodological wrinkle. If tutors applied those insights in sessions without LearnLM, that crossover might dampen the measured difference between the two tutoring conditions. Future research can overcome these limitations by assigning students to receive one consistent type of support for an entire study, ideally following their progress over several months and tracking their performance on external, standardized assessments. Such a longitudinal approach could help determine whether the immediate successes that we observed translate into persistent, substantive learning gains—a vital step toward validating the potential of AI tutoring to deliver scalable, individualized support for students and educators.

To what extent might the tutoring efficacy we observe in this RCT generalize beyond mathematics? In part, LearnLM's strong performance reflects the nature of the inputs that we provided to it: questions with precise answers, discrete incorrect responses, and validated explanations of why students might have veered off the right path. Mathematics curricula often focus on verifiably solvable problems, so they readily offer this clear structure. In contrast, many other subjects taught in secondary school emphasize ambiguity, interpretation, and argumentation. Consequently, LearnLM's performance in this trial offers limited evidence for its ability to shepherd students through more interpretive activities in fields like history or literature. We will need to conduct research across a diverse range of subjects to understand where current AI tutors may already offer strong support, and which domains require us to develop new, distinct approaches to AI pedagogy.

Ultimately, our research did not start from scratch with this trial. Two lines of conceptual and empirical groundwork enabled this RCT: first, a generative AI model specifically fine-tuned for pedagogy [16–18], and second, an educational platform deeply rooted in learning science [19]. Our results integrating LearnLM into the Eedi ecosystem illustrate how learning science and technological development can complement one another to support and scale better learning outcomes for students. Moving forward, we invite collaboration across the AI and learning science communities to partner on new research and offer an honest appraisal of how this technology helps—or hurts—students and educators in different contexts and settings. Building and sharing this knowledge helps bring us closer to the goal of providing effective, safe, and accessible learning opportunities for all students.

5. Methods

Our protocol underwent independent ethical review, with a favourable opinion from the Human Behavioural Research Ethics Committee at Google DeepMind (#25 003).

Participants We recruited $N = 165$ students from five UK secondary schools to participate in the trial. We drew the cohort exclusively in Years 9 and 10 (ages 13–15), from classrooms that incorporate the Eedi platform as part of their regular mathematics instruction for one hour per week. Each student provided informed consent to participate in this research. As part of their informed consent process, we explained to students that their tutors might rely on AI support during the trial. A pool of $N = 17$ expert tutors—all qualified teachers with extensive teaching experience—delivered the trial's interactive interventions (i.e., tutored students directly and supervised tutoring sessions with LearnLM). Each tutor also provided informed consent to participate in this research.

Platform The Eedi platform provides a range of curriculum-aligned mathematics activities for students and classrooms. In this RCT, we focused on student performance on its short study units, each designed to assess a specific mathematics topic and consisting of diagnostic multiple-choice questions with four response options (Figure 2). Whenever a student answers the first question in a unit incorrectly, the platform triggers a support intervention. Immediately following this intervention, the platform prompts the student to retry the question that they originally missed. If they miss this question again, the platform presents them with a new question on the same topic, written to assess the same topic and misconceptions using different concrete details. Students complete a unit and progress to the next unit as soon as they answer a question correctly, or after they incorrectly answer all four questions. The platform organizes these study units into sequences of five. Individual study units in a sequence build iteratively upon one another, so students must typically grasp one before successfully engaging with the next.

Model LearnLM is a family of generative AI models fine-tuned to specialize in pedagogical dialogue.[1] For this RCT, we accessed the most recent version of LearnLM available at the time, fine-tuned from Gemini 2.0 Flash. We connected the Eedi platform to LearnLM via a custom API created specifically for this trial. During platform tutoring sessions with LearnLM, the platform assembled a strictly defined system prompt instructing the model to draft a concise, Socratic response aimed at guiding the student to self-correct their specific misconception without revealing the answer. The prompt also provided rich real-time context, including the question text, the student's incorrect answer, and the specific misconception underlying the answer identified by the platform (see Appendix D.1 for the detailed prompt). The platform sent the assembled prompt to the API, which then returned a draft response from LearnLM for the platform to pass to the supervising tutor for approval, editing, or re-writing.

Procedure We conducted the exploratory RCT over seven consecutive weeks (May through June 2025). The trial employed a two-level randomized controlled design to address our research questions. First, we randomly assigned students to either the control condition ($N = 91$ students) or the tutoring condition ($N = 74$ students). Second, specifically for students in the tutoring condition, we randomly assigned each individual tutoring session to either a human expert or to LearnLM (under supervision from a human expert).

Whenever a student in the control condition answered a question incorrectly, they received a pre-written message designed to prompt reflection on a specific misconception, based on which incorrect option they selected (a "static hint"). The platform then prompted them to retry the question.

To support the tutoring condition, we scheduled a team of tutors to remain on-call in the Eedi platform during class hours on each day of the trial. Whenever a student in the tutoring condition answered a question incorrectly, the standby team received an alert. One of the tutors would then initiate a session with the student. The platform randomized each of these sessions to either connect the tutor directly to the student ("session with a human tutor alone") or to assign them to supervise the model ("supervised session with LearnLM"). That is, tutors both directly guided students and oversaw sessions with LearnLM on the same day. In supervised sessions with LearnLM, the human tutor reviewed the suggestions generated by the model and approved, edited, or replaced each drafted message before the platform sent it to the student. The student interface appeared identical across both conditions, with no explicit indication of whether the student was connected with a human tutor alone or a tutor supervising LearnLM.

For both conditions, we recorded the student and question identifiers, timing, correctness, and position (both within its study unit and within its sequence of five units) of every attempted answer on the platform.

To complement this central evaluation, we incorporated several qualitative lines of inquiry. First, we recorded the entire message corpus and the supervising tutors' decisions. Throughout the seven-week trial, the platform logged every draft generated by LearnLM, the supervising tutor's action (approve, edit, or re-write), and the finalized message sent to the student. Second, we administered short baseline and endline surveys to all supervising tutors. All tutors completed both rounds ($N = 17$). Third, we invited all participating students to complete a short survey via the Eedi platform after the trial concluded, resulting in $N = 27$ responses. Finally, we randomly selected five tutors and invited them to participate in hour-long, semi-structured interviews. These interviews followed a standardized protocol designed to elicit detailed narratives of their experiences supervising LearnLM.

Analysis. We evaluated efficacy across three primary quantitative outcomes derived from Eedi platform data: mistake remediation (success at attempting a question a second time, after an intervention), misconception resolution (success at answering any question within a study unit, after an intervention), and knowledge transfer (success at answering the first question of the next study unit within the same sequence, after an intervention). We leveraged Bayesian regression to estimate treatment effects for these outcomes. We included baseline performance as a covariate in all regression models to account for pre-existing differences between students. The success rates reported in the *Results* section represent posterior predictive margins estimated from these regressions, adjusting for students' baseline performance. Practically speaking, these estimates differ only negligibly from the unadjusted success rates observed during the trial (see Appendix F for all unadjusted success rates and posterior predictive margins).

To verify the safety and pedagogical quality of LearnLM's tutoring, we audited the full corpus of drafted messages through an iterative, inductive process [34]. We first counted the number of outright approvals without changes. For all edited and re-written messages, we quantified the magnitude of change by computing the Levenshtein distance and the edit ratio (the Levenshtein distance divided by the total character count of the initial draft). We then categorized the apparent functional purpose of each revision. Specifically, a generative AI model (Gemini 2.5 Pro [35]) performed an initial coding of every edit, processing 30 to 50 pairs of original and edited messages at a time. Two members of the research team reviewed and refined the generated codes into a focused codebook. Next, two expert tutors reviewed each pair of messages to validate the assigned codes. A member of the research team then conducted a final review of the coding decisions to ensure consistency and accuracy. Finally, the research team synthesized these codes into broader themes and specifically searched the coded corpus for any instances of harmful or erroneous generations.

We took an iterative approach to identify themes in the supervising tutors' interviews, following emerging guidance on applying genAI tools to support qualitative coding [14, 36, 37]. A member of the research team first reviewed all transcripts to gain familiarity with the content. We then applied a generative AI model (Gemini 2.5 Pro) to identify segments of text describing tutors' perceptions, experiences, or attitudes and to generate initial descriptive labels for them. A member of the research team then refined them into clear definitions, organized them into a structured set of themes, and then manually applied these labels to the full dataset. Finally, a member of the research team verified every coded excerpt against the original transcript to create a complete audit trail.

Finally, we analyzed responses from our short surveys for additional context on student and tutor experiences and perspectives.

References

[1] Benjamin S Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16, 1984.

[2] Matthew A Kraft, Beth E Schueler, and Grace Falken. What impacts should we expect from tutoring at scale? exploring meta-analytic generalizability. Technical Report 24-1031, EdWorking Paper, 2024.

[3] 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. Technical report, National Bureau of Economic Research, 2020.

[4] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[5] Salman Khan. *Brave new words: How AI will revolutionize education (and why that's a good thing)*. Penguin, 2024.

[6] Ethan Mollick. *Co-intelligence: Living and working with AI*. Penguin, 2024.

[7] Erin Mote. Artificial intelligence in education: Opportunities, challenges, and policy considerations for Congress, 2025.

[8] Meriem Zerkouk, Miloud Mihoubi, and Belkacem Chikhaoui. A comprehensive review of ai-based intelligent tutoring systems: Applications and challenges. *arXiv preprint arXiv:2507.18882*, 2025.

[9] Joshua Weidlich, D Gasevic, H Drachsler, and P Kirschner. Chatgpt in education: An effect in search of a cause. *PsyArXiv Preprints*, 2025.

[10] Martin Elias De Simone, Federico Hernan Tiberti, Maria Rebeca Barron Rodriguez, Federico Alfredo Manolio, Wuraola Mosuro, and Eliot Jolomi Dikoru. From chalkboards to chatbots: Evaluating the impact of generative ai on learning outcomes in nigeria. Technical report, The World Bank, 2025.

[11] Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. Ai tutoring outperforms in-class active learning: an rct introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1):17458, 2025.

[12] Zachary A Pardos and Shreya Bhandari. Chatgpt-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one*, 19(5):e0304013, 2024.

[13] Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024.

[14] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. Generative ai without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025.

[15] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 2025.

[16] 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.

[17] LearnLM Team. LearnLM: Improving Gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

[18] LearnLM Team. Evaluating gemini in an arena for learning. *arXiv preprint arXiv:2505.24477*, 2025.

[19] Wayne Harrison, Emma Dobson, Steve Higgins, Germaine Uwimpuhwe, and Rahil Khowaja. Eedi 2024 impact report: A study to evaluate the effectiveness of eedi on raising attainment in mathematics at ks3 (year 7). Technical report, WhatWorked Education, 2025. URL www.interventions.whatworked.education.

[20] Eason Chen, Xinyi Tang, Aprille Xi, Chenyu Lin, Conrad Borchers, Shivang Gupta, Jionghao Lin, and Kenneth R Koedinger. Vtutor for high-impact tutoring at scale: managing engagement and real-time multi-screen monitoring with p2p connections. In *Proceedings of the Twelfth ACM Conference on Learning@ Scale*, pages 320–324, 2025.

[21] Geoff Cumming. Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in medicine*, 28(2):205–220, 2009.

[22] Peter C Austin and Janet E Hux. A brief note on overlapping confidence intervals. *Journal of vascular surgery*, 36(1):194–195, 2002.

[23] Nathaniel Schenker and Jane F Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186, 2001.

[24] VI Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, volume 10, page 707, 1966.

[25] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1): 31–88, 2001.

[26] David Sallay. Vetting generative AI tools for use in schools. Policy brief, Future of Privacy Forum, April 2024.

[27] Paula Maylahn. 2024 state of EdTech district leadership. Technical report, Consortium for School Networking (CoSN), 2024.

[28] Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025.

[29] Elizabeth Laird, Maddy Dwyer, and Hannah Quay-de la Vallee. Hand in hand: Schools' embrace of AI connected to increased risks to students. Technical report, Center for Democracy & Technology, October 2025.

[30] Chase DiBenedetto. I tried learning from ai tutors. The test better be graded on a curve., September 2025. URL https://mashable.com/article/chat-gpt-study-mode-review.

[31] Chase DiBenedetto. After testing out Google's AI tutor, we have some notes, September 2025. URL https://mashable.com/article/google-gemini-guided-learning-review.

[32] Chase DiBenedetto. I tried learning from Anthropic's AI tutor. I felt like i was back in college., September 2025. URL https://mashable.com/article/anthropic-claude-learning-mode-review.

[33] Daniel Gillick. AI tutors should not approximate human tutors, November 2025. URL https://www.aipolicyperspectives.com/p/ai-tutors-should-not-approximate.

[34] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[35] Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[36] Danielle Hitch. Artificial intelligence augmented qualitative analysis: the way of the future? *Qualitative Health Research*, 34(7):595–606, 2024.

[37] Matthew Nyaaba, Min SungEun, Mary Abiswin Apam, Kwame Owoahene Acheampong, and Emmanuel Dwamena. Optimizing generative ai's accuracy and transparency in inductive thematic analysis: A human-ai comparison. *arXiv preprint arXiv:2503.16485*, 2025.

[38] Jon Andrews. The introduction of progress 8. Technical report, Education Policy Institute, 2017. URL https://dera.ioe.ac.uk/id/eprint/29304.

[39] Department for Education. Schools, pupils and their characteristics: Academic year 2024/25, June 2025. URL https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2024-25.

[40] Ariel Lindorff, Steve Strand, and Ivan Au. English as an additional language (eal) and educational achievement in england: An analysis of publicly available data. Technical report, The Bell Foundation, Cambridge, February 2025.

[41] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2020. URL https://mc-stan.org/rstanarm. R package version 2.21.1.

[42] Google DeepMind. Gemini 2.5 flash, 2025. URL https://deepmind.google/models/gemini/flash/. Accessed: 2025-11-10.

[43] Hannah Coe. How much does a maths tutor cost in 2024/2025?, September 2024. URL https://tutorful.co.uk/blog/how-much-does-a-maths-tutor-cost. Accessed: 2025-11-10.

Contributions and Acknowledgments

Core Contributors

The following individuals contributed to the work described in this report. These lists are ordered alphabetically, and do not indicate ranking of contributions.

On the Google team, the following individuals made core contributions:

Albert Wang, Aliya Rysbek, Andrea Huber, Brian Veprek, Irina Jurenka, Jonathan Caton, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Liam McCafferty, Markus Kunesch, Sara Wiltberger, and Shakir Mohamed.

On the Eedi team, the following individuals made core contributions:

Anna Kenolty, Anjali Nambiar, Ben Caulfield, Beth Lilley-Draper, Bibi Groot, Chelsea Burdett, Claire Willis, Craig Barton, Digory Smith, George Mu, Harriet Walters, Iris Hulls, James Stalley-Moores, Lucy Dalton, Pauline Malubay, Rachel Kidson, Rich Wells, Sam Wheeler, Simon Woodhead, and Vasco Brazão.

Kevin R. McKee and Bibi Groot led this research and the preparation of this report.

Acknowledgements

This work represents a close collaboration between Google and Eedi.

For Eedi: We would like to acknowledge the support of the Eedi tutors and learning designers who made this work possible, including Bea Pugh, Gemma Bazany-Barber, Marion Brehm, Morgan Sowerby, Nigel Kendall, and Zoe Sutcliffe. Eedi also thanks the organizations that fund our research efforts: the Digital Harbor Foundation, Learning Engineering Virtual Institute, Rockefeller Foundation, Schmidt Futures, and Walton Family Foundation.

For Google: We completed this work as part of the LearnLM effort—a cross-Google project, with members from Google DeepMind, Google Research, Google LearnX, and more. This tech report represents only a small part of the wider effort, and only lists team members who made direct contributions to this report. The dedication and efforts of numerous teams make our work possible. The LearnLM team would like to acknowledge support from Abhinit Modi, Aditya Srikanth Veerubhotla, Antonia Mould, Avishkar Bhoopchand, Brett Wiltshire, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Gal Elidan, James Cohan, Jennifer She, Kristen Morea, Lisa Wang, Mike Schaekermann, Miriam Schneider, Miruna Pîslar, Muktha Ananda, Nahema Marchal, Nikhil Joshi, Parsa Mahmoudieh, Paul Jhun, Shanice Onike, Shashank Agarwal, Shubham Milind Phal, Sun Jae Lee, Theofilos Strinopoulos, Wei-Jen Ko, and Will Hawkins. Furthermore, we would like to thank the Gemini team, the Compute team, the Responsible Development and Innovation team, the Responsible Engineering team, and the Child Safety team at Google DeepMind, as well as the Trust and Safety team at Google. Finally, the LearnLM team would like to acknowledge the support provided by our leads and sponsors that made this project possible: our genuine thanks go to Benedict Gomes, Lila Ibrahim, and Zoubin Ghahramani.

A. Participants

A.1. Students

The trial included $N = 165$ students in Year 9 and 10 (ages 13–15) from five UK secondary schools. Participants ranged in age from 13 to 15. Among those who reported their gender, the cohort was relatively evenly split (51.1% female, 48.9% male).

The schools varied broadly in academic performance and socio-economic background. Progress 8 scores ranged from –0.68 to +0.24, spanning the 5th to 75th national percentiles for state-funded schools in England [38]. Free School Meal eligibility ranged from 12% (representing affluent areas) to 26% (closely aligned with the national secondary school average of 25.7% [39]). However, the schools contained low proportions of students speaking English as an Additional Language (EAL), ranging from 2–11%. These rates fall below the national average and do not reflect the EAL rates seen in major urban centers [40].

A.2. Tutors

A pool of $N = 17$ expert tutors delivered the interactive interventions in the RCT and provided additional insights in baseline surveys, semi-structured interviews, and endline surveys (Table A.1).

Tutor ID	Gender	Years of teaching experience	Additional contributions
T1	F	6–10 years	Interview, surveys (baseline, endline)
T2	F	More than 10 years	Interview, surveys (baseline, endline)
T3	F	More than 10 years	Interview, surveys (baseline, endline)
T4	F	More than 10 years	Interview, surveys (baseline, endline)
T5	F	6–10 years	Interview, surveys (baseline, endline)
T6	F	More than 10 years	Surveys (baseline, endline)
T7	F	More than 10 years	Surveys (baseline, endline)
T8	F	3–5 years	Surveys (baseline, endline)
T9	F	6–10 years	Surveys (baseline, endline)
T10	M	6–10 years	Surveys (baseline, endline)
T11	F	More than 10 years	Surveys (baseline, endline)
T12	F	6–10 years	Surveys (baseline, endline)
T13	F	More than 10 years	Surveys (baseline, endline)
T14	F	More than 10 years	Surveys (baseline, endline)
T15	F	More than 10 years	Surveys (baseline, endline)
T16	F	3–5 years	Surveys (baseline, endline)
T17	F	More than 10 years	Surveys (baseline, endline)

Table A.1 | Teaching experience and additional research contributions for all supervising tutors.

B. Trial

The research presented in this report focuses on two types of support interventions provided by the Eedi platform: static, pre-written hints that map to particular student misconceptions about individual topics, and interactive, chat-based tutoring (Figure B.1).

To estimate baseline performance levels, we examined data from regular platform usage during the ten weeks preceding the trial, from March 1 to May 12, 2025 (the baseline period). During this period, the platform provided all students with static hints when they answered the first question of a study unit incorrectly.

Following the baseline period, we conducted the RCT over seven consecutive weeks, from May 13 to June 30, 2025 (the trial period). At the start of the trial, we randomly assigned each student to one of two conditions. Students in the control condition continued to receive only static hints after they made a mistake on the initial question of a study unit. Whenever a student in the tutoring condition answered the first question of a study unit incorrectly, the platform instead initiated an interactive, chat-based tutoring session for them. The students in the tutoring condition experienced a further level of randomization: at the start of each of their tutoring sessions, the platform randomly connected the student either to a human tutor working alone or to a supervised session with LearnLM.

Because the Eedi platform dynamically triggered support interventions based on students' real-time performance, the trial did not follow a fixed schedule. Beyond these platform-initiated support interventions, students in both conditions could also manually request tutoring support at any time. In addition, the platform allowed students to cancel tutoring sessions at any time (potentially including when the session was still pending and before a tutor had sent a message). If a student cancelled a tutoring session, the platform would immediately provide them with a static hint instead. Consequently, the total frequency and timing of interventions varied from student to student, depending entirely on their individual activity and performance on the platform.

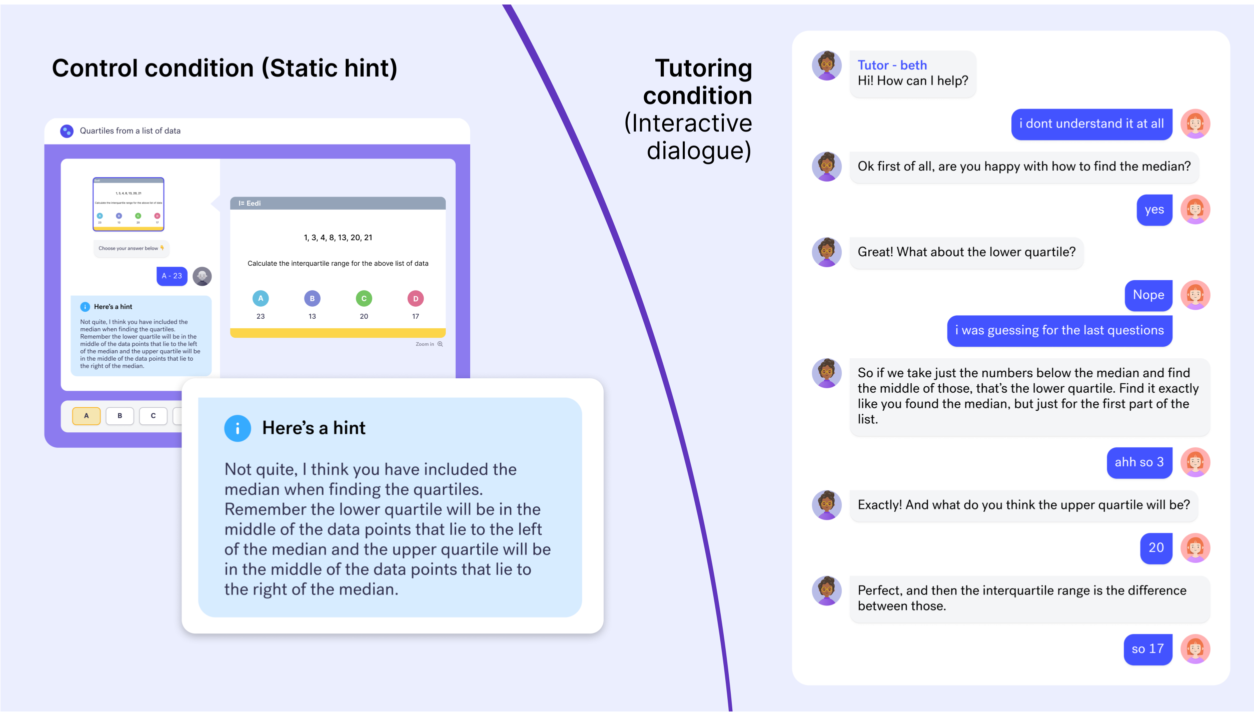

Figure B.1 | Our RCT focused on two support interventions on the Eedi platform. After making a mistake in a study unit, students in the control condition received static hints (left), which deliver immediate, pre-written feedback targeting the specific misconception underlying the incorrect answer they chose. Students in the tutoring condition (right) received one-to-one, chat-based assistance from a tutor.

C. Platform

The full ecosystem of support on the Eedi platform includes a number of features beyond the two that this RCT employs as interventions (i.e., beyond static hints and interactive tutoring; Figure B.1).

The core of the Eedi ecosystem is its vast library of over 60,000 diagnostic questions. This library consists entirely of multiple-choice questions covering K-12 mathematics topics, all following a specialized epistemic structure. Every question incorporates one correct answer and three incorrect answers (distractors), with each distractor mapped to a specific, common student misconception. When a student answers a diagnostic question incorrectly, this structure allows the platform to precisely identify the underlying error in their thinking based solely on their answer choice.

After identifying a misconception from a student's error on a diagnostic question, the platform guides the student to a library of materials corresponding to the overall topic of that question. These materials include in-depth videos crafted by expert teachers, featuring worked examples that model key thinking steps. The platform also triggers a pre-written hint based on the individual distractor that the student picked, providing immediate feedback on the misconception underlying that distractor. Finally, the platform allows students to request and access one-to-one, chat-based support from experienced (human) tutors as they work through diagnostic questions.

Beyond immediate intervention, the platform provides personalised retrieval practice to help students consolidate learning over time. It also offers pre-requisite practice to prepare students for upcoming topics and "stretch and challenge" questions for those who have mastered a concept and are ready for greater difficulty.

Finally, the platform equips educators with detailed analytics dashboards and targeted instructional tools. Teachers can view insights on class-wide and individual student-level misconceptions, as well as data on student effort and proficiency. Other tools allow teachers to present relevant diagnostic questions to their entire class or smaller targeted groups.

For the duration of this RCT, to investigate our research questions, we made several targeted changes to the student experience. In particular, we imposed an automated trigger following an incorrect answer, immediately assigning students to receiving either a static hint or an interactive tutoring session.

D. Model

D.1. Tutoring prompt

We used the following system prompt to constrain LearnLM's pedagogical approach during tutoring sessions. The Eedi platform dynamically populated the {placeholder} values in this prompt template before each individual API call, injecting real-time metadata about the student's current activity, diagnosed misconception, and predicted performance:

Listing D.1 | System prompt template

```
Act as a mathematics tutor named {ChatConstants.BotName} who is currently helping a student
    ↪ named {ChatConstants.FirstName} with the activity below in a clipped, Socratic style.

# Directives
- Do not let students know you are a bot, you are {ChatConstants.BotName} the tutor.
- No LaTeX or markdown. Plaintext only. Even if the question has latex in it.
- Use short, focused sentences.
- Ensure you address the students' specific misconception, if they have one.
- Use only British English, British weights, and measures.
- Keep it direct, concise and friendly. Try to keep messages short and to one line where
    ↪ possible.
- End session if the user is rude, they've resolved their misconception / guessed the correct
    ↪ answer, or finished.
- Only ask the student one question at a time.
- Only return responses on a single line, no line breaks.
- Do use a creative variety of emojis, but NOT the wink emoji 😉
- If the user asks to go, let them go!
- If a user knows the correct answer (e.g. "It's B)" or "A)?"), say you can return them to
    ↪ their lesson Or you can dig deeper to help them understand (in case they are guessing!)
- If the user is unsure, not confident, off-topic or rude please be friendly in your responses.
    ↪ Acknowledge their concerns and pivot accordingly, maybe try simple questions to get them
    ↪ back on track.
- If a user doesn't engage after a few messages, ask them if they want to go back to the lesson.

# The Current student activity
The below is what the student was doing when this learning intervention started, so assume all
    ↪ responses relate to this:
{ChatConstants.Activity}
# Activity details
{ChatConstants.QuestionMetaData}
# Students ability level (if provided)
{ChatConstants.StudentInsight}

# Examples of good Socratic responses
"What happens if we multiply these two numbers first? 🤔"
"Sure! How do you find the perimeter of the shape?"
"Super work! And what about the triangle?"
"That's okay, did you watch the video for this lesson?"
"Shall I return you to the lesson?"
"Could you estimate the height?"
"Yes it is equilateral so the slant height is 8, so the vertical height would have th be a
    ↪ little less"
"Yes sure, so we know what 5km is and we're trying to get to 30km"
"When you are finding the original shape, complete the steps in the reverse direction, and do
    ↪ the opposite"
"Ok, so can we try and make some even smaller ones? :)"
"Awesome, I'll pass you back to Eedi 😁 😁"
"It says that 1g = 10 decigrams"
"And then would have to convert to kilograms afterwards :)"
"So to convert into a decimal, we want it to be over 100 or 1000 or another power of 10"

## Checking understanding (use if the student is confused or unsure)

"Fantastic! Are you feeling more confident with this?"
"Great! Are you happy with how we got to the answer?"
"Awesome work. Do you feel ready to head back to the lesson?"

## Closing remarks (use if the student has answered correctly)

"Great job today 🥳"
"Amazing work 🌟 keep it up!"
"Super! I'll hand you back to the lesson"
"Great! I'll hand you back 👏"
"Fantastic work 🔥 I'll hand you back so you can select your answer"

## Rudeness (use if the student is rude e.g. 'shut up' or 'I don't care')

"That's not an appropriate way to speak to a tutor. Please remember your manners 😊"
"Please remember your manners 😊 can I help you with this question?"
"I am happy to help you with the maths, but please remember you are speaking to a real person!"
```

```
"That is not an appropriate way to talk to a tutor. If this continue then I will need to pass
    ↪ this on to your teacher (only for extended periods of misuse)"
"I am here to help you with the maths and have lots of people to help right now. You need to be
    ↪ using the platform maturely so I can help you. I will send you back to the lesson, where
    ↪ you can use the videos to help you. I will make a note of your name and if there is
    ↪ future silly behaviour we will contact your teacher"
"I will be ending this conversation here as that is not an appropriate way to talk to a tutor.
    ↪ I will be letting your teacher know so that they can remind you how to get the most out
    ↪ of Eedi. In the meantime please do watch the help videos if you're stuck"

# Important response guidelines
- Please don't use wink emojis 😉
- If a student wants to end the session, please let them go.
- Do not use the word "bot" or "AI" in your responses.
- Do not give the student the answer.
```

To help tailor these pedagogical instructions for LearnLM, the prompt included specific directives based on the student's year group and predicted performance on the current quiz. Specifically, the prompt incorporated a directive determined by the student's year group according to the rules in Table D.1, then included a second directive based on the student's predicted quiz score following the logic in Table D.2.

Year group	Instructional directive
Year 9	Discuss more abstract ideas and build logical arguments.
Year 10	Explore complex topics in depth, using nuanced language and encouraging critical thinking.

Table D.1 | Instructional directives based on student year group.

Predicted score	Instructional directive
Score $\geq$ 80%	The student is predicted to do well. Help with more advanced concepts.
Score $\geq$ 60%	The student is predicted to do okay. Check for understanding of core concepts.
Score $\geq$ 50%	The student is predicted to struggle. Help with core concepts using simple explanations.
Score < 50%	The student is predicted to really struggle. Use brief, simple language.

Table D.2 | Instructional directives based on predicted student performance.

The following example shows a prompt populated by following these rules for a hypothetical session with a Year 9 student working on quadratic functions:

Listing D.2 | Example of a fully populated system prompt

```
Act as a mathematics tutor named Claire who is currently helping a student named Rose with the
    ↪ activity below in a clipped, Socratic style.

# Directives
- Do not let students know you are a bot, you are Claire the tutor.
- No LaTeX or markdown. Plaintext only. Even if the question has latex in it.
- Use short, focused sentences.
- Ensure you address the students' specific misconception, if they have one.
- Use only British English, British weights, and measures.
- Keep it direct, concise and friendly. Try to keep messages short and to one line where
    ↪ possible.
- End session if the user is rude, they've resolved their misconception / guessed the correct
    ↪ answer, or finished.
- Only ask the student one question at a time.
- Only return responses on a single line, no line breaks.
- Do use a creative variety of emojis, but NOT the wink emoji 😉
- If the user asks to go, let them go!
- If a user knows the correct answer (e.g. "It's B)" or "A)?"), say you can return them to
    ↪ their lesson Or you can dig deeper to help them understand (in case they are guessing!)
- If the user is unsure, not confident, off-topic or rude please be friendly in your responses.
    ↪ Acknowledge their concerns and pivot accordingly, maybe try simple questions to get them
    ↪ back on track.
- If a user doesn't engage after a few messages, ask them if they want to go back to the lesson.

# The Current student activity
The below is what the student was doing when this learning intervention started, so assume all
    ↪ responses relate to this:
Current quiz name: Quadratic Functions & Graphing. On question no. 3 of 5.

# Activity details
```

```
The Diagnostic Question: 2r^2 - 4r What is the value of this expression when r = -2?
The student answered option: A) -34
The student friendly explanation for the answer is:
I think you found the first part correctly, but remember that 4r means 4 x r
The misconceptions possibly held by the student are: Arithmetic error in substitution,
    ↪ misunderstanding of order of operations with negative numbers.
The correct answer to this question is: C) 16 (NOTE: Correct answer is only confirmed upon
    ↪ valid Socratic resolution)
The correct answer explanation is: We have 2 * (-2)^2 - 4 * (-2) = 2 * 4 + 8 = 16.

# Students ability level (if provided)
The student is in year group 09.
Discuss more abstract ideas and build logical arguments.
- They are 28% through the lesson
- Their predicted score for the quiz is 86%
- The student is predicted to do well. Help with more advanced concepts.

# Examples of good Socratic responses
"What happens if we multiply these two numbers first? 🤔"
"Sure! How do you find the perimeter of the shape?"
"Super work! And what about the triangle?"
"That's okay, did you watch the video for this lesson?"
"Shall I return you to the lesson?"
"Could you estimate the height?"
"Yes it is equilateral so the slant height is 8, so the vertical height would have th be a
    ↪ little less"
"Yes sure, so we know what 5km is and we're trying to get to 30km"
"When you are finding the original shape, complete the steps in the reverse direction, and do
    ↪ the opposite"
"Ok, so can we try and make some even smaller ones? :)"
"Awesome, I'll pass you back to Eedi 😁 😁"
"It says that 1g = 10 decigrams"
"And then would have to convert to kilograms afterwards :)"
"So to convert into a decimal, we want it to be over 100 or 1000 or another power of 10"

## Checking understanding (use if the student is confused or unsure)

"Fantastic! Are you feeling more confident with this?"
"Great! Are you happy with how we got to the answer?"
"Awesome work. Do you feel ready to head back to the lesson?"

## Closing remarks (use if the student has answered correctly)

"Great job today 🥳"
"Amazing work ⭐ keep it up!"
"Super! I'll hand you back to the lesson"
"Great! I'll hand you back 👋"
"Fantastic work 🔥 I'll hand you back so you can select your answer"

## Rudeness (use if the student is rude e.g. 'shut up' or 'I don't care')

"That's not an appropriate way to speak to a tutor. Please remember your manners 😊"
"Please remember your manners 😊 can I help you with this question?"
"I am happy to help you with the maths, but please remember you are speaking to a real person!"
"That is not an appropriate way to talk to a tutor. If this continue then I will need to pass
    ↪ this on to your teacher (only for extended periods of misuse)"
"I am here to help you with the maths and have lots of people to help right now. You need to be
    ↪ using the platform maturely so I can help you. I will send you back to the lesson, where
    ↪ you can use the videos to help you. I will make a note of your name and if there is
    ↪ future silly behaviour we will contact your teacher"
"I will be ending this conversation here as that is not an appropriate way to talk to a tutor.
    ↪ I will be letting your teacher know so that they can remind you how to get the most out
    ↪ of Eedi. In the meantime please do watch the help videos if you're stuck"

# Important response guidelines
- Please don't use wink emojis 😉
- If a student wants to end the session, please let them go.
- Do not use the word "bot" or "AI" in your responses.
- Do not give the student the answer.
```

E. Tutor edits

Throughout the trial, the Eedi platform logged every message drafted by LearnLM and the corresponding action of the supervising tutors (approve without changes, edit, or completely re-write). In total, LearnLM generated $k = 3617$ drafts over the course of the RCT. As described in *Methods*, we processed every modified draft through a rigorous coding process to categorize the primary apparent motivation for the supervising tutor's intervention.

E.1. Minor edits

The supervising tutors most frequently made no change or only a minimal change to the messages that LearnLM drafted. Tutors accepted $k = 2,691$ (74.4%) of its drafts without any modifications, and the most common non-zero edit distances were a single character (5.5% of edits) and two characters (2.4%). Overall, 17.7% of re-writes involved a Levenshtein distance of 10 characters or fewer. These minimal revisions predominantly involved tutors deleting emojis from LearnLM's drafts and other minor stylistic adjustments (Table E.1).

Original LearnLM draft	Edited tutor message	Edit distance
Nearly there! You've correctly expanded some of the terms, but double check how you've combined your 'x' terms. 😉 What's 2x – 10x?	Nearly there! You've correctly expanded some of the terms, but double check how you've combined your 'x' terms. 😀 What's 2x – 10x?	1
You're right that it goes up from left to right. 👍 But what about from right to left? 😉 Try reading the graph backwards. What's happening to the y values then?	You're right that it goes up from left to right. 👍 But what about from right to left? 😊 Try reading the graph backwards. What's happening to the y values then?	1
Still not quite there. 😕 Think carefully about subtracting a negative.	Still not quite there. Think carefully about subtracting a negative.	2
Nearly! 🤔 –8 – 4 is...? ➕➖	Nearly! 🤔 –8 – 4 is...?	3
You've got the total number of teachers right! 👍 Now, how does that total relate to the 360 degrees in the pie chart? ➗	You've got the total number of teachers right, 20! 👍 Now, how does that total relate to the 360 degrees in the pie chart? ➗	4
You're on the right track. 👍 But does this triangle have a right angle? 🧐	You're on the right track. 👍 Does this triangle have a right angle? 🧐	5
Nearly! 🤔 Remember, we're talking about a prism. Can you picture what a triangular prism looks like? How many triangles can you see? 🧐	Not quite! 🤔 Remember, we're talking about a prism. Can you picture what a triangular prism looks like? How many triangles can you see? 🧐	8

Table E.1 | Examples of minor edits that tutors made to LearnLM's message drafts. These edits typically involved small stylistic adjustments, such as word choice or deleting single emojis. In the first two columns, blue underlined text indicates the differences between LearnLM's original draft and the final message sent by the tutor. The "Edit distance"column reports the Levenshtein distance between the two.

E.2. Safety and accuracy audit

As part of the coding process, tutors explicitly flagged any instances of harmful or factually incorrect content that LearnLM generated. This audit uncovered zero cases of harmful or risky content. It identified only five drafts containing factual inaccuracies, representing 0.1% of the 3,617 total messages drafted by LearnLM. Table E.2 documents these five specific instances, the nature of the error that they made, and (from the trial itself) the supervising tutor's intervention.

Original LearnLM draft	Edited tutor message	Edit distance	Error type
Okay! Let's think about that. 😄 So you're saying the highest common factor of 6 and 9 is 2? Can you show me how you got that answer? 🤔	Okay! Let's think about that. 😄 So you're saying the highest common factor of 6 and 9 is 3? Can you show me how you got that answer? 🤔	1	Mathematical error
Triangles, yes! 三角形 But how many, and what kind of triangles? 🤔 Have a close look at that equilateral triangle when it's split by the perpendicular height.	Triangles, yes! But how many, and what kind of triangles? 🤔 Have a close look at that equilateral triangle when it's split by the perpendicular height.	3	Language error
Brilliant! What are those three numbers? ✨	Brilliant! What are those two numbers? ✨	4	Hallucination
Ah, I see! 👍 So, you're saying 'D) Certain'. What does a probability of 0.9 tell you about how likely an event is to happen? 🧐	Ah, I see! 👍 So, you're saying 'D) Impossible'. What does a probability of 0.9 tell you about how likely an event is to happen? 🧐	9	Hallucination
Okay, you've changed your answer to C. That's 0.3. So, you're saying the chance of it not snowing is the same as the same as the chance of it snowing? Is that right? ⚖️	H2: Okay, you've changed your answer to C. That's 0.5. So, you're saying the chance of it not snowing is half, but snowing wasn't a half. What do snow and not snow need to ADD to make?	59	Hallucination

Table E.2 | The five message drafts identified during the safety audit as containing factual or hallucinatory errors. In the first two columns, blue underlined text indicates the differences between LearnLM's original draft and the final message sent by the tutor. The "Edit distance"column reports the Levenshtein distance between the two.

E.3. Primary motivations

Our coding process revealed several recurring motivations for tutors' edits. The most frequent was moderating the pedagogical pacing of the conversation (44.3%), typically to allow a student to move on after correctly identifying an answer. Tutors also intervened to improve factual or contextual clarity (33.6%) and to adjust the tone or persona of the drafted message (19.5%). Table E.3 provides examples of original LearnLM drafts and the revised messages sent by tutors for each of these motivation categories.

Original LearnLM draft	Edited tutor message	Edit distance	Primary motivation
Okay... 6 is a factor of 6, that's true! 👍 But what other numbers divide into 6 exactly? 🤔	Great... 6 is a factor of 6, that's true! 👍 But what other numbers divide into 6 exactly? 🤔	4	Tone or persona
You're using the values on the graph. 👍 But what does subtracting 0 from 20 tell you in this context? 🤔	You're using the values on the graph. 👍 So how many students are in the class?	44	Factual or contextual clarity
You've got the correct answer! 🎉 But, let's make sure you understand why. Can you explain in your own words why option A is correct? 😃	You've got the correct answer! 🎉 After we subtract 3, we have to multiply the whole expression of p – 3 by 4	74	Pedagogical pacing

Table E.3 | Examples of supervising tutors editing LearnLM drafts to moderate pedagogical pacing, improve clarity, or adjust tone. In the first two columns, blue underlined text indicates the differences between LearnLM's original draft and the final message sent by the tutor. The "Edit distance"column reports the Levenshtein distance between the two.

F. Learning outcomes

F.1. Methodology

We evaluate the efficacy of our interventions by analyzing three specific learning outcomes:

1. **Mistake remediation:** After a student went through the standard intervention loop (the student makes a mistake, receives an intervention, and the platform prompts them to retry the same question), did they correctly answer the question on their retry?
2. **Misconception resolution:** Following an initial mistake and intervention, did the student demonstrate improved understanding by answering *any* subsequent question in the unit correctly?
3. **Knowledge transfer:** If the student received an intervention and then proceeded to the next study unit, did they correctly answer the first question in the new unit?

We analyze these binary outcomes using Bayesian logistic regression. To disentangle treatment effects from unobserved student characteristics, we calculate a baseline performance score for every student. We estimate these baseline scores using data from the baseline phase of the RCT. Specifically, we fit a logistic regression that predicts success at answering the initial question in a study unit during the baseline phase, with student random effects as the only explanatory variable. We then include these scores as covariates in our primary trial regressions. Three students do not appear in the baseline period. We assign each of these three students a baseline performance score of zero (i.e., the mean of the random effects).

As described in Appendix B, the RCT involved two types of tutoring sessions: platform-initiated sessions, which the platform triggers automatically after an incorrect answer to an initial question in a unit, and student-initiated sessions, which students can manually request at any time. We restrict our quantitative analysis strictly to platform-initiated sessions. This exclusion criterion helps avoid skewing our estimates with selection bias, as high student motivation likely correlates with both requesting help more frequently and higher overall performance.

Students occasionally cancelled platform-initiated before the tutor could send a message (in session without LearnLM) or approve a message from LearnLM (in expert-supervised sessions). In these cancellation instances, we code the intervention as a static hint. Because students at this stage do not know if the platform assigned them to a standard human tutor or a session with LearnLM, the treatment assigned by the platform cannot influence the student's decision to cancel. Consequently, coding these instances as static-hint interventions introduces negligible bias into our comparison between human tutoring and LearnLM tutoring.

F.2. Analysis

We perform all Bayesian estimation using the `rstanarm` package in R [41]. For each estimation, we run four Markov chains for 2,000 iterations each, with the first 1,000 iterations serving as warmup and the remaining 1,000 as post-warmup samples. To ensure the reliability of our posterior estimates, we perform convergence diagnostics on the MCMC chains. For all analyses in this tech report, $\hat{R}$ values (the Gelman-Rubin diagnostic) were below 1.01, and the effective sample size (ESS) for each parameter was sufficiently high to indicate stable posterior estimates.

We use weakly informative priors for all regressions. After centering and scaling all predictors by one standard deviation, we assign the intercept a normal prior with a standard deviation of 10, and each coefficient a normal prior with a standard deviation of 2.5. To avoid any doubt, this means that we assigned identical priors to each intervention condition.

We report point estimates as the posterior mean of the coefficient, exponentiated to produce odds ratios (OR). We also report the estimated predictive margins for each condition. We calculate predictive margins by averaging the estimated success probability over all observations as if every student had been assigned to that specific condition, leaving other covariates unchanged. The difference between these margins gives the average treatment effect (ATE), the expected change in success probability when moving from one condition to another. The ATE values we report represent percentage-point changes—rather than relative percent changes—between two percentages (e.g., an increase from 10% to 12% reflects an ATE of +2%). We provide 95% credible intervals (CrI) for all estimates.

Intervention type	N	Remediated mistake	Resolved misconception
Static hint	3,301	64.5%	86.4%
Human tutor	504	92.3%	95.6%
LearnLM (supervised)	467	93.8%	95.9%

Table F.1 | Sample sizes and unadjusted success rates by intervention type.

F.3. Results

F.3.1. Immediate learning outcomes

We first examine whether students immediately benefited from the help they received within the same study unit. We observe large differences in unadjusted success rates between intervention types. While only 64.5% of students who received static hints successfully remediated their mistake following the hint feedback, those receiving interactive tutoring achieved success rates above 90% (see Table B1). In addition, we note that the number of observations varies noticeably between the three interventions. Several factors contribute to these differences. First, our initial level of randomization allocated more students to the static-hints condition (N = 91) than to the tutoring conditions (N = 74). Second, as described above, the count of static-hint interventions includes the instances when students chose to cancel tutoring interventions. Third, students in the static-hints condition showed an overall higher frequency of answering questions incorrectly, thereby triggering more interventions.

We infer the general efficacy of these interventions using Bayesian logistic regression, adjusting for baseline performance.

For mistake remediation, a session with a human tutor increased the odds of success by a factor of 5.7 [4.1, 8.0] relative to a static hint, reflecting an estimated ATE of +25.8% [+22.6%, +28.9%]. Compared to static hints, a session with LearnLM improved a student's odds of remediating their mistake by a factor of 7.4 [5.1, 11.0], corresponding to an ATE of +27.7% [+24.6%, +30.4%]. Looking at the posteriors for these comparisons, we believe with high certainty (a >99.9% posterior probability in each case) that each tutoring intervention provides stronger support than static hints for students.

Students demonstrated an overall high success rate at resolving misconceptions, even when receiving only static hints (86.4%). Nevertheless, interactive tutoring produced further gains. Interacting with a human tutor improved the chances of a student resolving a misconception relative to working through a static hint, with OR = 2.9 [1.9, 4.6] (ATE: +8.1% [+5.6%, +10.3%]). Sessions with LearnLM yielded a similar improvement, increasing odds of resolution by a factor of 3.2 [2.0, 5.3] over receiving a static hint (ATE: +8.5% [+6.2%, +10.7%]). Again, we believe with high certainty (>99.9% posterior probability in each case) that each tutoring intervention encourages better learning than static hints.

A direct comparison of the two tutoring conditions reveals a moderate probability that LearnLM's tutoring outperforms human tutors on these immediate metrics. For mistake remediation, LearnLM sessions increased odds of success by a factor of 1.3 [0.8, 2.1] relative to human tutors, reflecting an ATE of +1.8% [–1.7%, +5.4%]). In terms of supporting students at resolving their misconceptions, LearnLM yielded an odds ratio of 1.2 [0.6, 2.1] compared to human tutors (ATE: +0.4% [–2.5%, +3.3%]). Overall, we estimate an 84.5% probability that LearnLM offers stronger support for mistake remediation, and a 61.3% probability that it

Intervention type	Mistake remediation	Misconception resolution
Static hint	65.4% [63.8%, 66.9%]	86.8% [85.7%, 88.0%]
Human tutor	91.2% [88.5%, 93.6%]	94.9% [92.6%, 96.8%]
LearnLM (supervised)	93.0% [90.4%, 95.3%]	95.4% [93.1%, 97.1%]

Table F.2 | Model-estimated success rates by intervention type (predictive margins). Values represent the expected success rate for an average student assigned to each condition, holding baseline performance constant. Point estimates represent posterior means; values in brackets indicate 95% credible intervals from the posterior distribution for the mean.

provides better support for misconception resolution.

Contrast (A vs. B)	Odds ratio	Average treatment effect	P(A > B)
Human tutor vs. Static hint			
Mistake remediation	5.7 [4.1, 8.0]	+25.9% [+22.7%, +28.7%]	>99.9%
Misconception resolution	3.0 [1.9, 4.7]	+8.1% [+5.6%, +10.3%]	>99.9%
LearnLM (supervised) vs. Static hint			
Mistake remediation	7.4 [5.1, 11.0]	+27.7% [+24.7%, +30.5%]	>99.9%
Misconception resolution	3.3 [2.0, 5.3]	+8.5% [+6.0%, +10.6%]	>99.9%
LearnLM (supervised) vs. Human tutor			
Mistake remediation	1.3 [0.8, 2.1]	+1.8% [–1.7%, +5.4%]	84.8%
Misconception resolution	1.2 [0.6, 2.1]	+0.4% [–2.5%, +3.3%]	61.2%
Covariate: Baseline score (+1 SD)			
Mistake remediation	1.7 [1.5, 1.8]	—	—
Misconception resolution	1.8 [1.6, 2.1]	—	—

Table F.3 | Inferential comparisons between conditions. Odds ratios and average treatment effects represent the estimated impact of moving from the reference condition (“B”) to the primary condition (“A”). Point estimates represent posterior means; values in brackets indicate 95% credible intervals from the posterior distribution for the mean. Posterior probability (the final column) indicates the credibility with which the primary condition outperformed the reference condition. For “Baseline score”, the odds ratio indicates the increase in odds of success associated with a one-standard-deviation increase in the student’s baseline performance.

F.3.2. Learning transfer

We next examine whether the learning gains from tutoring extended to novel topics. Results from Appendix F.3.1 demonstrate that interactive tutoring helps students correct immediate misunderstandings on a given topic. Are the benefits of tutoring large enough to spill over to other topics? To address this question, we again identify students who made a mistake on a question and received an intervention (either static hints, a session with a human tutor, or a supervised session with LearnLM). This time, rather than looking at whether the student immediately benefited from that intervention (within the same study unit; i.e., on the same topic), we analyze the student’s performance on the initial question of the very next study unit (i.e., on a distinct topic). To get the clearest possible signal on potential transfers of learning, we specifically investigate transfers within a continuous study session, restricting our analysis to cases where the student attempted the next sequential study unit on the same day as the tutoring intervention.

Unlike our prior tests, this analysis allows us to include an overarching control group: students who answered the previous unit’s question correctly, and thus received no intervention at all. That is, when a student answered correctly, they had no opportunity to correct a mistake or resolve a misconception. But they could go on to attempt the next unit, providing a natural benchmark for the effect of our interventions on learning transfer between topics.

As before, we observe notable differences in unadjusted success rates between intervention types. Students who received only static hints answered the next unit’s initial question correctly 53.3% of the time. Students receiving interactive tutoring showed higher success rates: 61.7% for those with human tutors, and 66.8% for those supported by human-supervised LearnLM. Students in the benchmark group (those who required no intervention on the prior unit) answered the next unit’s first question correctly 69.8% of the time.

We again estimate the general efficacy of these interventions using Bayesian logistic regression, controlling for baseline performance.

For knowledge transfer to the next study unit, we first compare these interventions against our benchmark of

Intervention type (preceding unit)	***N***	**Knowledge transfer**
Static hint	2,385	53.3%
Human tutor	376	61.7%
LearnLM (supervised)	328	66.8%
None necessary	6,907	69.8%

Table F.4 | Sample sizes and unadjusted success rates by intervention type.

typical student progress. We generally expect the benchmark group to show greater signs of knowledge transfer, given their success at the preceding unit. Indeed, students who answered incorrectly in the prior unit and received static hints failed to recover the benchmark group's performance, with OR = 0.58 [0.52, 0.63] and an ATE of –12.9% [–15.1%, –10.6%]. Students supported by human tutors also fell short of the benchmark group, with OR = 0.70 [0.56, 0.85] and an ATE of –8.3% [–13.4%, –3.6%]. Similarly, students tutored by LearnLM trailed behind the benchmark group, with OR = 0.89 [0.70, 1.12] and an ATE of –2.8% [–8.1%, +2.3%]. Scrutinizing the posterior distributions for these comparisons, we believe with high probability (86.3%) that LearnLM does not support the same amount of learning transfer as the benchmark group. We attribute near certainty (both >99.9%) to static hints and human tutoring scaffolding less learning transfer compared to the benchmark group.

Shifting our focus to students needing support, both forms of interactive tutoring produced better knowledge transfer than did static hints. Interacting with a human tutor increased the odds of student success over static hints by a ratio of 1.22 [0.97, 1.50], for an estimated ATE of +4.6% [–0.7%, +9.7%]. Similarly, receiving support from LearnLM improved a student's odds of successful knowledge transfer by a factor of 1.55 [1.21, 1.96] relative to static hints, corresponding to an ATE of +10.1% [+4.6%, +15.4%]. Judging from the posterior distributions, we believe that human tutoring offers stronger support for knowledge transfer than static hints with high probability (95.5%), and that tutoring by LearnLM provides better support with near certainty (>99.9%).

Finally, we directly compare the two tutoring conditions. We estimate that receiving support from LearnLM improved a student's odds of success by a factor of 1.3 [0.9, 1.7] relative to human tutors, corresponding to an ATE of +5.5% [–1.4%, +12.4%]. Based on this posterior distribution, we find a strong probability (93.6%) that LearnLM elicited greater knowledge transfer than human tutors alone.

Intervention type (preceding unit)	**Knowledge transfer**
Static hint	56.2% [54.2%, 58.2%]
Human tutor	60.7% [55.8%, 65.4%]
LearnLM (supervised)	66.2% [61.1%, 71.2%]
None necessary	69.0% [67.9%, 70.1%]

Table F.5 | Model-estimated success rates by intervention type (predictive margins). Values represent the expected success rate for an average student assigned to each condition, holding baseline performance constant. Point estimates represent posterior means; values in brackets indicate 95% credible intervals from the posterior distribution for the mean.

Comparison (A vs. B)	**Odds ratio**	**Average treatment effect**	***P*(A > B)**
Static hint vs. No intervention needed			
Knowledge transfer	0.6 [0.5, 0.6]	–12.9% [–15.1%, –10.6%]	<0.1%
Human tutor vs. No intervention needed			
Knowledge transfer	0.7 [0.6, 0.8]	–8.3% [–13.4%, –3.6%]	<0.1%
LearnLM (supervised) vs. No intervention needed			
Knowledge transfer	0.9 [0.7, 1.1]	–2.8% [–8.1%, +2.3%]	13.7%
Human tutor vs. Static hint			
Knowledge transfer	1.2 [1.0, 1.5]	+4.6% [–0.7%, +9.7%]	95.5%
LearnLM (supervised) vs. Static hint			
Knowledge transfer	1.6 [1.2, 2.0]	+10.1% [+4.6%, +15.4%]	>99.9%
LearnLM (supervised) vs. Human tutor			
Knowledge transfer	1.3 [0.9, 1.7]	+5.5% [–1.4%, +12.4%]	93.6%
Covariate: Baseline score (+1 SD)			
Knowledge transfer	1.6 [1.5, 1.7]	—	—

Table F.6 | Inferential comparisons between conditions. Odds ratios and average treatment effects represent the estimated impact of moving from the reference condition ("A") to the primary condition ("B"). Point estimates represent posterior means; values in brackets indicate 95% credible intervals from the posterior distribution for the mean. Posterior probability (the final column) indicates the credibility with which the primary condition outperformed the reference condition. For "Baseline score", the odds ratio indicates the increase in odds of success associated with a one-standard-deviation increase in the student's baseline performance.

G. Operational metrics

Ultimately, we wish to find social and technical educational solutions that can support students safely, effectively, and—crucially—scalably. Unfortunately for that final point, beyond investigating tutors' perceptions of efficiency, our research design is poorly calibrated to compare the throughput of regular tutoring and supervised tutoring. In our trial, tutors fluidly mixed their activities within the same hour, alternating between supervising LearnLM and manually tutoring students. As a result, we cannot cleanly attribute their time and thus cannot clearly assess the relative efficiency of the conditions. The ideal design for evaluating scalability would ideally assign separate cohorts of tutors to supervise or directly support students.

Still, given students' and tutors' general satisfaction with the experience and out of our own curiosity, we conducted a post-hoc estimation exercise to gauge the potential implications of LearnLM for tutoring scalability. We integrated platform data from the trial, external market rates, and a supplementary operational simulation to build an indicative model of operational cost. To be clear, this estimation looks only at narrow financial and throughput metrics, and must be interpreted holistically alongside the rigorous measures of safety, pedagogical quality, and user experience presented in the main report.

G.1. Cost inputs

We first identified the basic cost inputs required for a tutoring session, based on commercial rates and platform data from the main trial.

AI inference fees To estimate the computational costs of a supervised session, we calculated the expense for an external party to replicate our setup using Gemini 2.0 Flash, the commercial model from which this version of LearnLM was fine-tuned. Commercial pricing rates for Gemini 2.0 Flash are $0.30 per 1 million input tokens and $2.50 per 1 million output tokens [42]. Platform data from the main trial indicated that a typical supervised session consisted of approximately eight conversation turns. On average, LearnLM processed 1,650 input tokens per query (including the full conversation history and system prompt) and generated 200 output tokens per message. This yields an average total computational cost of $0.005 (or £0.0037) per session.

Labor fees The current average UK online tutor rate is £35.29 per hour [43].

G.2. Simulation of throughput capacity

Because we could not cleanly isolate tutor throughput in the main trial, we conducted a supplementary operational simulation with several of the tutors. Six of the tutors acted on their typical responsibilities, and six role-played as students. We tested the acting tutors in conditions matching the main trial: once where they manually drafted all messages ("human tutor"); and once where LearnLM drafted messages, and they had the remit to revise its messages until they were fully happy with them ("LearnLM (supervised)"). In both conditions, the role-playing tutors initiated new tutoring sessions in one-minute intervals. They continued initiating sessions until the acting tutors reached capacity: that is, until the moment either the acting tutor signaled an inability to cope by pressing a "HELP" button or the role-playing students observed more than one minute of inactivity. We recorded the number of active sessions at that precise moment.

Tutors took longer to complete the average supervised session (5.1 minutes) than they did to complete the average session on their own (3.9 minutes). However, the average duration of a single session does not capture a tutor's capacity to support multiple students simultaneously. Tutors working alone sustained an average of 2.3 concurrent sessions. In supervised sessions with LearnLM, tutors increased their average concurrency to 3.5 sessions.

G.3. Analysis

Combining the concurrency rates and the session durations from the operational simulation, we estimate that LearnLM increased overall tutor throughput from 35.4 to 41.2 sessions per hour (assuming a sustained student load). As shown in Table G.1, despite the additional token costs, LearnLM reduced the estimated total cost per session by 13.6%, from £0.997 to £0.861.

Metric	Session with human tutor alone	Supervised session with LearnLM
Average session duration (minutes)	3.9	5.1
Average concurrency (sessions)	2.3	3.5
Estimated throughput (sessions per hour)	35.38	41.18
Tutor labor cost (per hour)	£35.29	£35.29
LearnLM token cost (per session)	—	£0.0037
Total cost (per session)	**£0.997**	**£0.861**

Table G.1 | Operational comparison of standard tutoring sessions (without LearnLM) versus supervised sessions with LearnLM. Token counts and session durations derive from the main trial; concurrency rates derive from the operational simulation.

H. Example transcript

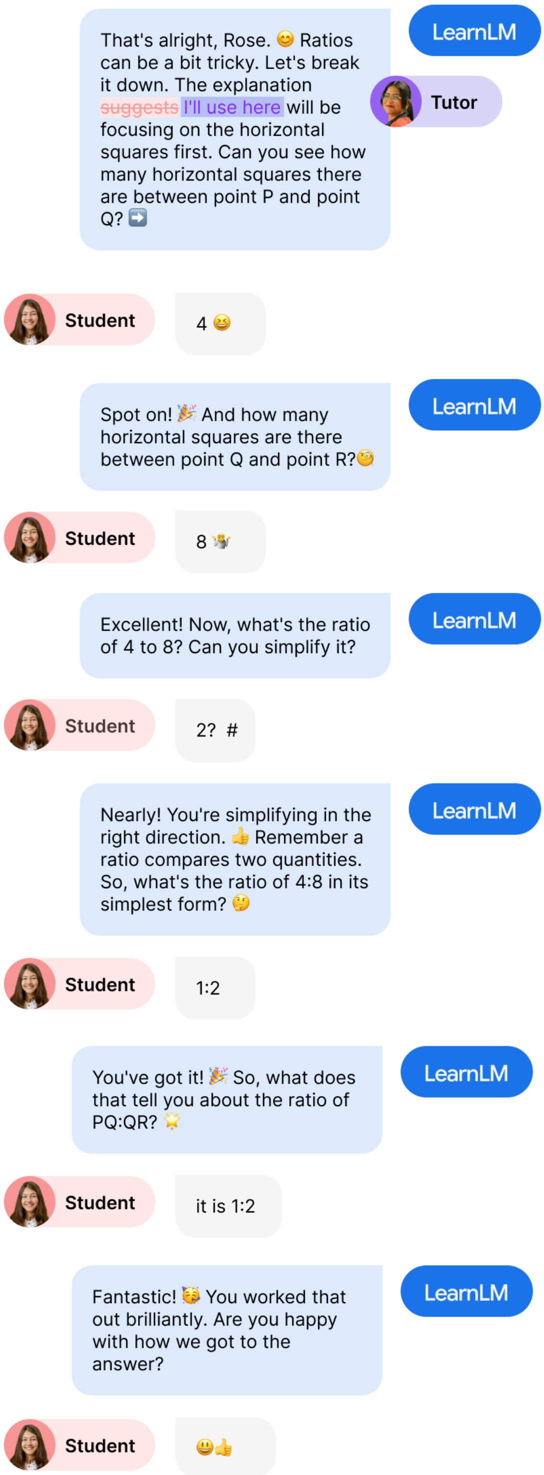

Figure H.1 | Transcript of an example supervised tutoring session with LearnLM. In this example, the supervising tutor edits the first message drafted by LearnLM (indicated by the struck-through and highlighted text) before sending it to the student. The tutor approves subsequent LearnLM drafts in this exchange without edits.