Measuring the impact of AI tutoring, from near transfer to standardised tests: An expanded RCT in UK mathematics classrooms

LearnLM Team, Google & Eedi

Randomised controlled trials of AI in classrooms remain rare, and few attempt to trace the effects of AI pedagogy from individual interactions to sustained learning. Earlier this year, we conducted a pre-registered RCT with the Eedi mathematics platform and Gemini across 10 state secondary schools in the UK. Over 10 weeks of weekly mathematics lessons, the trial compared the impact of four types of support for $N = 1179$ students: standard remediation on the Eedi platform, including curated videos and expert-written hints to resolve student misconceptions; expert human tutoring; AI tutoring, guided by a specialised pedagogical prompt; and AI tutoring, guided by the pedagogical prompt and personalised context about the student's learning history. To better evaluate the breadth of learning within these conditions, we measured student progress at three different levels: proximate measures of understanding within the Eedi platform; an intermediate, midtrial quiz; and standardised Star Maths assessments, administered and scored externally at baseline and endline.

During the trial, expert tutors reviewed every message drafted by Gemini, either approving, regenerating, or rewriting them before delivery. Out of 94,385 drafts, the supervising tutors ended up approving 94.7% of messages without any changes. Retrospective audits found mathematical errors in fewer than three in every 1,000 AI messages—matching the accuracy of expert tutors—and flagged potential tone issues in only two drafts, neither of which reached a student.

On proximate learning measures and the intermediate quiz, students receiving either variant of AI tutoring performed similarly to those tutored by expert humans, and substantially better than those receiving standard remediation. Unexpectedly, we did not detect such clear benefits on Star Maths: students demonstrated similar gains after either 10 weeks of standard remediation or 10 weeks of AI tutoring. We identify several tentative explanations for the incongruity between our narrow and broad measures of learning. For one, tutored students spent more time resolving misconceptions in depth, which decreased the breadth of material they covered—particularly early in the trial. In addition, exploratory analyses found the effects of tutoring on Star Maths to be more favourable for girls, with boys not yet seeing the same broad benefits. Looking forward, we argue that future research should investigate the timescale and conditions required for the immediate gains of AI tutoring to accumulate and emerge on broader assessments.

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

1. Introduction

Last year, we conducted an exploratory randomised controlled trial (RCT) of AI tutoring in mathematics classrooms across five UK secondary schools [1]. The trial tested whether embedding a generative AI system fine-tuned for pedagogy [2] into the Eedi mathematics platform[1] could safely and reliably provide academic support to students. The results were promising [3]. Expert human tutors judged roughly three in every four AI-drafted messages to meet their own pedagogical standards. Students receiving AI tutoring performed at

[1] https://www.eedi.com/

Corresponding author(s): learnlm-tech-report@google.com
© 2026 Google. All rights reserved

least as well as those tutored by expert tutors on all measures of learning in the Eedi platform, and even showed signs of stronger near transfer to subsequent topics.

These findings demonstrated that AI tutoring can operate safely and reliably in real classrooms, but stopped short of testing whether AI tutoring can drive cumulative gains in student learning. Because we intended the exploratory RCT primarily to evaluate the safety and pedagogical quality of AI tutoring, we targeted a smaller scale for the study, randomised support across sessions rather than across students, and measured learning through immediate, in-platform metrics. Whether immediate improvements from AI tutoring translate into sustained, broader learning remains an open question across the field. Rigorous classroom trials of AI in education are still rare, and most experiments evaluate brief exposures or narrow, task-specific outcomes [4–7] (see Appendix A for a review).

Answering this question requires a deeper and longer view of student progress in the classroom. In practice, this means running a longer trial, assigning each student to a single condition throughout the trial to allow gains to accumulate, and measuring progress on general subject assessments. For example, independent researchers previously conducted a classroom RCT over a two-year period to properly assess the impact of the Eedi platform on students' standardised test scores [8].

Earlier this year, we took a step toward this longer-term goal, conducting a scaled, pre-registered RCT of AI tutoring with $N = 1179$ students across 10 UK state secondary schools over 10 weeks of classwork. We embedded Gemini models[2] into the Eedi platform under the supervision of expert tutors in order to evaluate whether AI tutoring remains safe and pedagogically reliable at scale, and whether it improves student learning—traced from proximate in-platform measures, through an intermediate curriculum quiz, to a broad, independently administered standardised test.

2. Trial design

The Eedi platform supports classroom mathematics instruction by organising typical curriculum topics into short study units, each consisting of a pair of multiple-choice "diagnostic questions" [9]. In these diagnostic questions, every incorrect option (i.e., distractor [10]) corresponds to a specific, common error in mathematical reasoning. A student's selection of a distractor thus identifies a potential misconception to correct, turning a mistake into a clear signal for intervention and instruction (see also [11, 12]). The Eedi team worked with experienced educators to craft each diagnostic question and its constituent misconceptions. They also created the platform's standard support for addressing those misconceptions: instructional videos, fluency exercises, and a set of short, tailored hints, each mapped to a distractor and written to help students recognise and correct their mistakes. Students begin each study unit by attempting an initial diagnostic question (the "check-in"). A correct answer advances them directly to the next unit, whereas an incorrect answer triggers targeted support for their misconception, followed by a second diagnostic question on the same topic (the "check-out"). Teachers assign sequences of these study units to match the curricula for their classes. Students then work through the assigned units at their own pace during their weekly mathematics lessons. Prior research provides strong empirical support for this approach. Over the two-year, independent study that we mentioned above, students working through the platform's diagnostic questions and standard remediation achieved learning gains equivalent to roughly four additional months of progress, relative to peers who were assigned not to use the platform [8].

For this new trial, we partnered with 10 schools that incorporate the Eedi platform as a regular part of their mathematics curriculum. We recruited $N = 1179$ students in Years 8 to 10 across these schools, randomly assigning each student to one of four experimental conditions for 10 weeks of classroom activity, with students using the platform during their regular mathematics lessons for roughly 30 to 50 minutes per week. The full trial window included a week-long school holiday and the endline testing period, so the research spanned 12 calendar weeks overall.

The four conditions varied the support that students received after answering a diagnostic question incorrectly (Figure 1). As a control condition, the Eedi platform provided its *standard remediation* after a student made a mistake (i.e., self-guided instructional videos, fluency exercises, and a tailored hint for their specific misconception). For these students, a mistake did not trigger a session with a tutor. The other three conditions replaced the self-guided resources comprising standard remediation with a live, one-to-one online chat to

[2] https://ai.google.dev/gemini-api

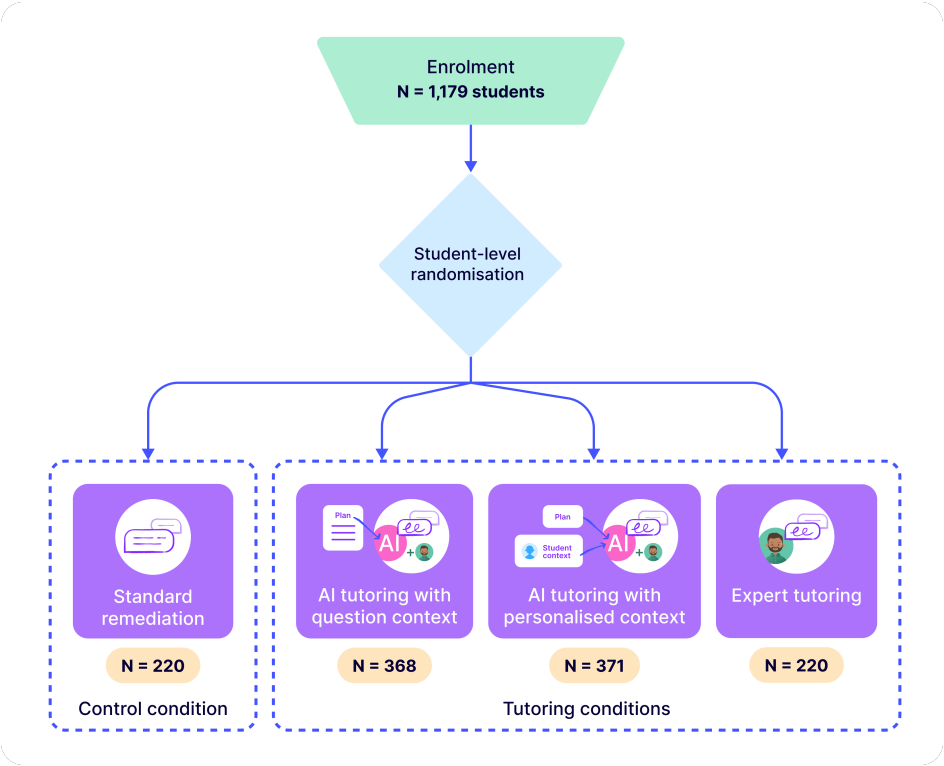

Figure 1 | We designed this 10-week classroom RCT to evaluate the safety, pedagogy, and learning impact of AI tutoring. We randomly assigned $N = 1179$ students to receive one of four types of support when they answered the initial diagnostic question in a study unit incorrectly: the platform's instructional videos, fluency exercises, and tailored hints (*standard remediation*); or a one-to-one online chat with a remote mathematics educator (*expert tutoring*), with Gemini guided by the current question and misconception (*AI tutoring with question context*), or with Gemini guided by additional context on the student's learning history (*AI tutoring with personalised context*).

guide the student through their misconception (i.e., "problem-based tutoring" [13, 14]).[3] Students receiving *expert tutoring* joined a chat with a remote mathematics educator [15–17]. These tutors wrote every message from scratch, without any AI assistance. In the next condition, *AI tutoring with question context*, an AI model drafted each message for the student, guided by a specialised pedagogical prompt that incorporated the specific diagnostic question at hand, as well as the misconception behind the student's incorrect answer. The same model also drafted the messages for the *AI tutoring with personalised context* condition, which we designed to include the same pedagogical prompt, diagnostic question, and misconception, plus additional context on the student's recent learning history (drawn from their journey through past study units and earlier tutoring chats; cf. [18]).

We built both AI tutoring conditions around Gemini, drawing on its fine-tuning for pedagogical use cases [2, 19]. In particular, Google DeepMind fine-tunes the Gemini model family to follow specific teaching instructions—rather than simply providing direct answers—across a wide range of learning interactions.[4] As a result, Gemini models can readily be prompted to apply pedagogical principles such as encouraging students to actively work through problems, breaking complex material into manageable pieces, and adapting explanations to each individual student. For this trial, we configured the Eedi platform to call Gemini via API to generate message drafts for the AI tutoring conversations. That is, all conversations took place inside the standard chat interface on the Eedi platform, whether the student interacted with an expert tutor or the AI tutor (Figure 2).

To ensure these conversations delivered effective instructional support, we designed the core pedagogical

[3]The Eedi platform also includes an optional "Ask for help" button that students can click to request a chat with a tutor, at any time. This trial focuses specifically on varying the support triggered automatically after mistakes, so we allowed students in all four conditions—including those receiving standard remediation—to use the button as normal throughout the trial (see *Methods* and Appendix B).

[4]https://cloud.google.com/solutions/learnlm

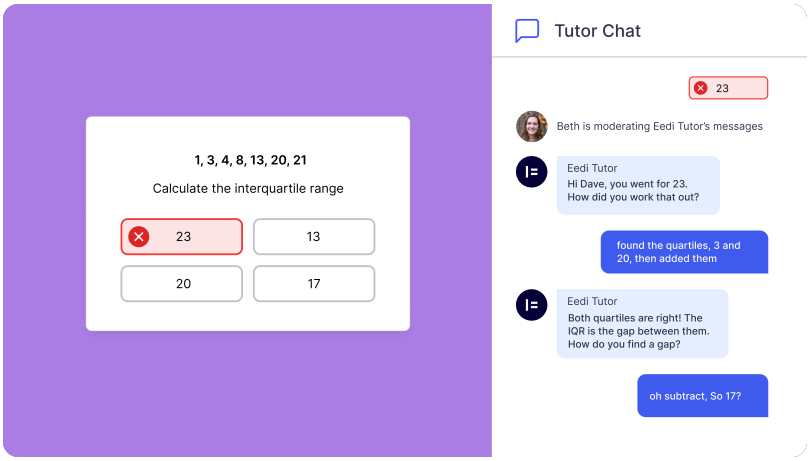

Figure 2 | All tutoring conversations took place inside the standard chat interface on the Eedi platform, whether the student interacted with an expert tutor or Gemini.

prompt for this trial around best practices in mathematics tutoring. This core prompt instructed Gemini to guide students through their misconceptions step by step with concise, scaffolded questions, without giving away the correct answer. It also instructed the model to tailor its teaching strategy to the type of mistake the student made—for instance, using a simpler example to expose a flawed method, or walking through a partially worked example to address a procedural error (see Appendix C). In both AI tutoring conditions, the platform appended the question context for each conversation to this core prompt: the diagnostic question, the answer options, the correct solution, and the educator-written explanation of the student's misconception. In the personalised condition, we designed the platform to additionally append a summary of the student's performance across past study units, recurring misconceptions, and earlier tutoring chats. Before drafting its first message to a student, Gemini used this context to generate a (private) teaching plan for the session, which the platform subsequently included in each prompt to Gemini to draft a message to the student. Given the rapid pace of AI development, we updated the tutor to call the best available Gemini models as new versions launched during the trial, from Gemini 3 Flash and 3.1 Pro to 3.5 Flash (see *Methods*).

We recruited $N = 58$ experienced mathematics educators to serve as expert tutors—most sourced directly from Eedi's tutoring team. Each tutor rotated between daily shifts tutoring students directly for the *expert tutoring* condition, without AI assistance, and separate shifts supervising Gemini for the AI tutoring conditions. During their supervisory shifts, tutors monitored Gemini's tutoring for safety and pedagogical quality. The platform required explicit approval from the tutor before sending any message to the student. In the AI tutoring conditions, for each draft generated by Gemini, the supervising tutor could approve the message without changes, regenerate it, or edit and rewrite it as needed.

For problem-based tutoring to build into broad mathematics attainment, its effects must carry through several stages. To start, each tutoring conversation needs to help the student better understand the topic at hand. Next, students must retain and consolidate these narrow learning gains over subsequent weeks. And ultimately, these individual sessions must span a wide range of topics to accumulate into general proficiency. To trace student progress across these stages, we measured learning at three corresponding levels (Figure 3). First, we tracked two proximate measures on the Eedi platform during weekly lessons: student accuracy on the second diagnostic question in a study unit immediately after receiving support (immediate mistake correction [21, 22]), and accuracy on the initial diagnostic question of the next study unit (near transfer [23, 24]). Second, as an intermediate measure, teachers administered a short quiz partway through the trial, covering the curriculum topics that each class had studied thus far. This quiz tested students' consolidation of the concepts they had practised over the preceding weeks (intermediate consolidation [25, 26]). Finally, as a broad measure, students completed the Renaissance Star Maths assessment [20] at baseline and endline. This independent, standardised test gauges students' general mathematical knowledge and skills. To avoid any doubts, students completed each of these outcome assessments by themselves, without any concurrent assistance from a tutor or AI. As a result, all of our learning outcomes reflect students' own proficiency.

Beyond students' mathematical proficiency, we also wanted to understand the perspectives and experiences

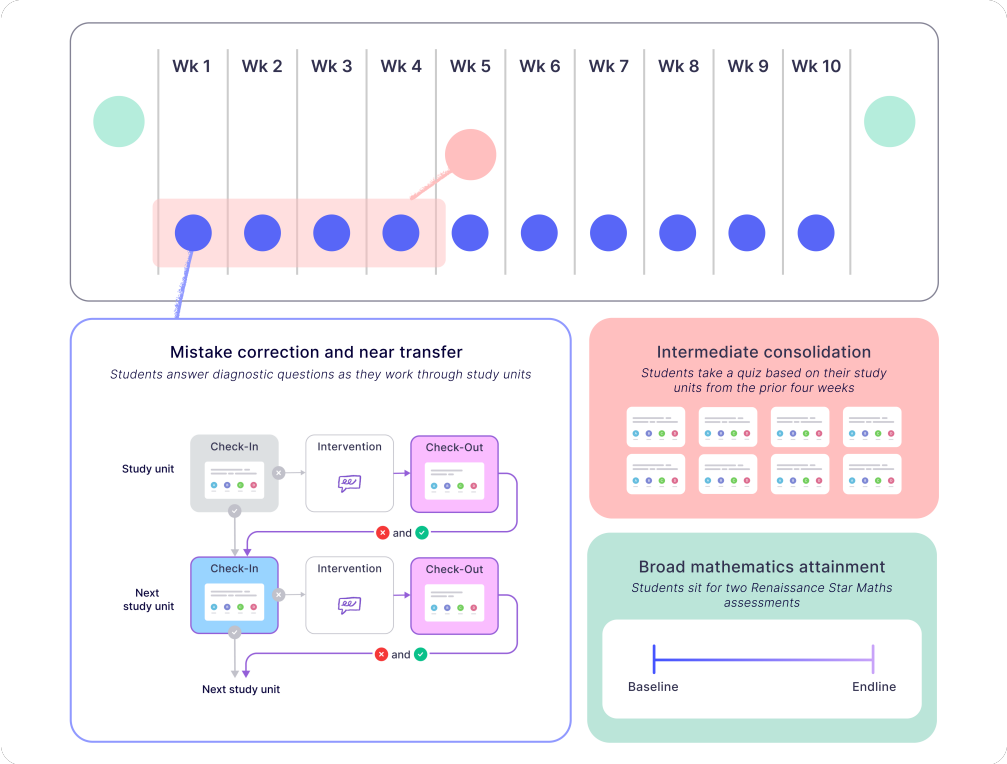

Figure 3 | To trace the effects of tutoring from individual interactions to broad mathematics attainment, we measured student learning at three levels. Specifically, we assessed two proximate measures on the Eedi platform (immediate mistake correction and near transfer), an intermediate midtrial curriculum quiz, and a broad standardised exam at baseline and endline (Renaissance Star Maths [20]).

of the students and tutors in our trial. These experiences can influence and shape students' academic progress—and matter as outcomes of their own, too. We therefore consulted students and tutors for their perspectives throughout the trial. We administered short surveys to students and tutors at baseline and endline. We also asked tutors to complete brief daily diaries after each of their shifts, and to share their reflections on the trial during a set of endline interviews.

An independent ethics committee reviewed and approved our trial protocol, including our approach to engaging with students, parents, and tutors, as well as our procedures for experts to supervise AI tutoring sessions (see *Methods*).

Together, this design allowed us to investigate five research questions:

RQ1: Was Gemini a safe, reliable, and pedagogically sound source of tutoring when deployed at scale under expert supervision?

RQ2: Did one-to-one tutoring (whether *expert tutoring* or *AI tutoring*) improve student learning—across proximate, intermediate, and broad measures—relative to *standard remediation* on the Eedi platform?

RQ3: For students receiving one-to-one tutoring, was supervised *AI tutoring* as effective for student learning as *expert tutoring*?

RQ4: Did providing Gemini with additional context on a student's learning history (*AI tutoring with personalised context*) improve tutoring and student learning relative to providing only the current question and misconception (*AI tutoring with question context*)?

RQ5: Over the trial, how did students experience tutoring, and how did expert tutors experience supervising Gemini?

We pre-registered the design, hypotheses, and analysis plan for this trial (AEARCTR-0018079 [27]). To answer our research questions, we adopted a Bayesian framework to evaluate both the magnitude and the credibility of our treatment effects. Bayesian approaches allow direct estimation of a posterior probability

distribution over possible effect sizes, providing a practical foundation for real-world deployment decisions. Across our analyses, we assigned identical, weakly informative priors to each condition so that the observed data would drive our estimates. From the resulting posterior distributions, we report 95% credible intervals for our estimates, alongside directional posterior probabilities (e.g., the probability that outcomes in one condition exceeded those in another). For complete experimental and analytical details, see *Methods* and Appendices B–H.

3. Results

3.1. Safety and pedagogical reliability

We first evaluated whether Gemini provided safe and pedagogically reliable tutoring when deployed at scale under expert supervision (**RQ1**). Across the trial, Gemini generated a total of 114,935 draft messages across the two AI tutoring conditions. Because the platform prompted Gemini to generate a fresh draft whenever a student sent another message, some drafts were superseded before the supervising tutor could review them. The supervising tutors thus reviewed and made decisions on a slightly smaller corpus of 94,385 drafts. They approved the vast majority of these without any changes: 89,406, or 94.7% of the non-superseded drafts. This approval rate substantially exceeded the rate observed in our 2025 exploratory trial, where tutors approved 74.4% of drafts unchanged—reflecting a 5.87-fold reduction in the odds of a draft requiring correction (95% credible interval of [4.01, 8.50]). Looking at whole conversations, 77.1% of AI tutoring sessions ran from start to finish without any revisions from the supervising tutor.

Across the remainder of the reviewed drafts (4,979 drafts, or 5.3%), supervising tutors either prompted Gemini to regenerate a response or rewrote the message themselves. When tutors edited a draft themselves, they typically rewrote most of the message rather than making minor tweaks, changing a median of 69% of the original text. To understand why tutors intervened, we retrospectively coded the apparent goal of each revision. Tutors most often stepped in to adjust style and tone (32.5%), refine the pedagogical scaffolding (28.6%), or manage student engagement (22.9%). Looking at the timing of revisions, we also see that tutors interceded most often near the end of a conversation. Managing student engagement accounted for 40% of late revisions, compared with 16% of early revisions. At the end of tutoring chats, these revisions most often reflected an attempt to regulate pedagogical pacing. That is, tutors interceded to allow students to move on, rather than continuing the line of guided questioning from Gemini (see Appendix D).

We next scrutinised whether this high approval rate reflected genuine draft quality. In particular, we wanted to rule out the possibility that supervising tutors simply rubber-stamped Gemini's drafts without carefully reviewing them. To evaluate draft quality independently of the supervising tutors' live decisions, we audited the delivered tutoring transcripts against a pedagogical rubric (see Appendix D). Across the AI tutoring conditions, the messages reaching students consistently followed our pedagogical guidelines. The AI tutor withheld direct answers until the student tried to work out the answer themselves in 98.8% of conversations, compared with 96.3% in the expert tutoring condition. Once the student and tutor resolved the student's original mistake, the AI tutor asked a follow-up question to check understanding in 91.3% of conversations, relative to 60.0% for expert tutors. We similarly saw no sign of tutors growing complacent as they became familiar with Gemini and with their supervising responsibilities: across the trial, weekly approval rates fell in a narrow range between 92.8% and 96.4%.

Finally, to assess accuracy and safety, we replicated the retrospective audit from our 2025 exploratory trial across all 114,935 generated drafts, including those superseded before review. This audit confirmed that Gemini rarely produced messages containing mathematical errors or potential tone issues. We flagged only six drafts across the entire corpus (<0.01%), a sharp reduction from the 2025 trial (0.14%). In two of these drafts, Gemini raised a potential tone issue by adding a winking emoji when declining a student's request to move on (Table 1, top two rows). The other four drafts contained minor mathematical errors in which Gemini misread a number in the question (Table 1, bottom four rows). None of these six drafts reached a student. Supervising tutors replaced five of them during live review, and a new student message superseded the sixth before the supervising tutor could review it.

Some mathematical errors, however, only become apparent in the full context of a conversation: a tutor might praise a student's incorrect answer, for example. We therefore conducted a second audit of the tutoring conversations for contextual errors, coding whether the tutor marked a wrong answer right, marked a right

Original Gemini draft	Issue	Explanation	Resolution
"Nice try, but I'm not letting you off the hook that easily! 😉 We are so close to the end. How many 3s go into 12?"	Tone issue	Uses a winking emoji that could read as overly familiar	Corrected
"I'm not letting you off that easily! 😉 Is 5^2 equal to 5×5 or 5×2?"	Tone issue	Uses a winking emoji that could read as overly familiar	Superseded
"No worries, we can do this! We need to check all the options. What is the area of Option A (4 cm by 4 cm)?"	Maths error	Describes Option A as 4 cm by 4 cm rather than 1 cm by 16 cm	Corrected
"On the first shape, the 10 cm side is on the left..."	Maths error	Places the 10 cm side on the left rather than on the bottom	Corrected
"Almost! If we start at 200 and take 3 jumps of 5, what is $200 + 5 + 5 + 5$?"	Maths error	Uses jumps of 5 rather than jumps of 10	Corrected
"Take another close look. Is the 5 cm side on the bottom, or is it on the left?"	Maths error	Places the 5 cm side on the bottom or the left rather than on the right	Corrected

Table 1 | Our initial retrospective audit of all 114,935 messages drafted by Gemini identified mathematical errors and potential tone issues in six drafts (<0.01%). None of the six drafts reached a student: supervising tutors corrected five during live review, and a new student message caused a new draft to supersede the sixth.

answer wrong, or made any other mathematical mistakes that specifically became clear *in situ* (see Appendix D). This audit revealed that Gemini and the expert tutors rarely made such errors. Gemini made a contextual error in 0.24% of its messages, relative to 0.18% for the expert tutors. Gemini and the expert tutors marked a wrong answer as correct at nearly the same rate (0.16% and 0.15%, respectively), while Gemini slightly more often marked a correct answer as wrong (0.08% and 0.03%, respectively), usually with an "Almost!" or "Not quite!" This audit also caught a small number of other mathematical mistakes that the message-by-message audit missed. In particular, we identified mistakes in seven messages delivered across AI tutoring conversations (0.01%) and in seven messages delivered across expert tutoring conversations (0.02%) that only became clear with conversational context.

3.2. Proximate and intermediate measures of learning

Next, we evaluated whether one-to-one tutoring improved student learning relative to standard remediation (**RQ2**), and whether AI tutoring was as effective as expert tutoring (**RQ3**). We began with our most proximate learning measure: whether students corrected their immediate mistake. On the Eedi platform, we capture this as student accuracy on the second diagnostic question in a study unit, after receiving support. Individual tutoring substantially improved this metric. When working with an expert tutor, students answered the second diagnostic question correctly 57.7% of the time [55.5%, 60.0%], compared with only 43.4% [41.5%, 45.3%] for students receiving standard remediation. AI tutoring proved just as effective: students corrected their mistake 56.9% [55.2%, 58.6%] of the time when receiving AI tutoring with question context, and 56.5% [54.8%, 58.2%] of the time when receiving AI tutoring with personalised context. Altogether, we attribute near certainty (each >99.9%) to expert tutoring and each form of AI tutoring outperforming standard remediation, and find that both forms of AI tutoring match the effectiveness of expert tutoring (Figure 4, left).

We also examined near transfer [23, 24]: the extent to which support on one problem helps students succeed when they move on to a new, related problem. Because study units on Eedi build sequentially on one another, we measured near transfer as student accuracy on the initial diagnostic question of the very next unit, after receiving support. As expected, one-to-one tutoring promoted substantially greater near transfer than standard remediation. Students who received standard remediation answered the next unit's initial diagnostic question correctly 50.5% [48.4%, 52.5%] of the time. By comparison, accuracy rose to 59.1% [56.9%, 61.3%] for students who worked with an expert tutor, 58.4% [56.8%, 60.0%] for those who received AI tutoring with question context, and 58.3% [56.7%, 59.9%] for those who received AI tutoring with personalised context. Here, too, we attribute near certainty (each >99.9%) to the advantage of expert tutoring and each form of AI

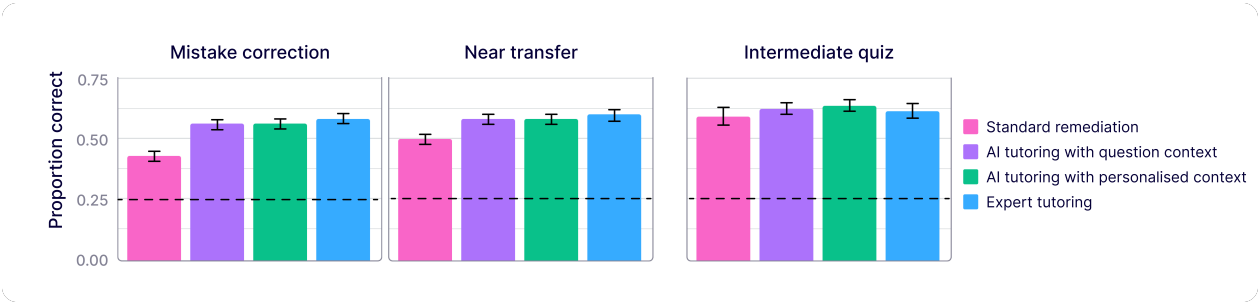

Figure 4 | One-to-one tutoring improved student performance across proximate and intermediate learning measures. (left, centre) For proximate in-platform outcomes, expert tutoring and both forms of AI tutoring promoted similar levels of mistake correction and near transfer. Students in all three tutoring conditions substantially outperformed students receiving standard remediation. (right) On the intermediate curriculum quiz, tutored students continued to score higher than students receiving standard remediation. Bar height and error bars indicate posterior estimates and 95% credible intervals, whereas dashed lines represent random chance.

tutoring over standard remediation. For near transfer, once again students tutored by Gemini matched the performance of students tutored by experts (Figure 4, centre).

To appraise students' consolidation of these gains across topics, we turned to our intermediate measure of learning. Partway through the trial, each teacher administered an eight-question quiz covering the curriculum topics that their class had studied over the preceding four weeks. Tutoring continued to improve student performance on this intermediate quiz. Students who received standard remediation answered 59.6% [56.4%, 62.8%] of the quiz questions correctly. By comparison, students answered 61.8% [58.5%, 64.8%] of the questions correctly when working with an expert tutor, 62.1% [59.7%, 64.4%] when receiving AI tutoring with question context, and 63.5% [61.2%, 65.9%] when receiving AI tutoring with personalised context. Altogether, we attribute a high credibility (97.1%) to AI tutoring with personalised context improving quiz performance over standard remediation, with more modest evidence of advantages for AI tutoring with question context (87.8%) and expert tutoring (81.3%) over standard remediation. Considering expert tutoring itself as a baseline, we estimate only a 57.3% probability of an advantage for AI tutoring with question context, compared with an 80.2% probability that students received more effective support from AI tutoring with personalised context (Figure 4, right; see also Appendix E for all comparisons).

We conducted one additional check on the persistence of learning gains over time. As part of its regular practice flow, the Eedi platform periodically brings back diagnostic questions from earlier study units to reinforce previous topics. In an exploratory analysis, we examined student accuracy when a diagnostic question returned more than seven days after a student had initially missed it and received support. Individual tutoring showed a clear advantage on these delayed questions. Students who received standard remediation answered the returning question correctly 47.6% [44.1%, 51.1%] of the time. In contrast, students answered correctly 58.9% [54.9%, 63.0%] of the time after working with an expert tutor, 59.9% [57.1%, 62.8%] after receiving AI tutoring with question context, and 59.0% [56.0%, 61.8%] after receiving AI tutoring with personalised context. Across these returning questions, we attribute near certainty (each >99.9%) to all three tutoring conditions outperforming standard remediation, and find that both forms of AI tutoring again matched the performance of expert tutoring (see Appendix E).

3.3. Broad mathematics attainment

Our next analysis focused on student performance on our broad measure of learning, the Renaissance Star Maths assessment. Star Maths is an independent, standardised test that many UK schools routinely administer to track student progress across national mathematics curricula [20, 28]. While individual tutoring produced clear gains on our proximate and intermediate measures, the picture shifted on this broader assessment (Figure 5). From baseline to endline, we did not see an overall advantage on Star Maths for tutoring over standard remediation. Students receiving AI tutoring with question context achieved outcomes almost identical to students receiving standard remediation, differing by only –0.01 standard deviations [–0.15, +0.14]. Students receiving AI

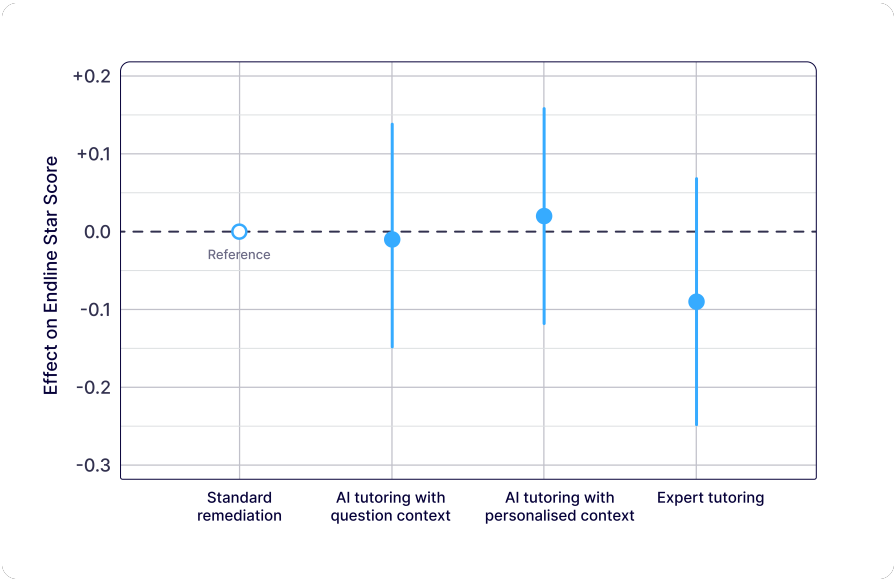

Figure 5 | Across the full sample, one-to-one tutoring did not improve endline Star Maths scores relative to standard remediation. Students receiving either form of AI tutoring achieved baseline-to-endline gains nearly identical to those of students receiving standard remediation. Unexpectedly, students in standard remediation and both AI tutoring conditions outperformed students working with an expert tutor. Points and error bars indicate posterior estimates and 95% credible intervals.

tutoring with personalised context scored similarly, with an effect of +0.02 SD relative to standard remediation [–0.12, +0.16]. More surprisingly, expert tutoring lagged behind all three other conditions. During the trial, working with an expert tutor produced an effect of –0.09 SD [–0.25, +0.07]. Comparing across conditions, we estimate an 87.1% probability that expert tutoring underperformed standard remediation, an 88.5% probability that it underperformed AI tutoring with question context, and a 94.4% probability that it underperformed AI tutoring with personalised context (see Appendix E for all model specifications). On the surface, then, our pre-registered analysis of Star Maths points to two surprising patterns: no overall effect of AI tutoring, and an apparent dip for expert tutoring.

These patterns seem to contradict the positive effects of tutoring that we observed on our proximate and intermediate measures. They also run counter to an abundance of prior research showing that individual tutoring improves students' standardised test scores [15]. We thus conducted several exploratory analyses to try to make better sense of these results.

Unlike our proximate and intermediate measures, Star Maths tests a wide range of topics across the curriculum, so our students' scores depend in part on the amount of material they cover over the trial. Within a fixed weekly lesson, students who review a short hint after an error might move through study units more rapidly than students who engage in a multi-turn tutoring conversation [29]. To investigate this possibility, we analysed data from the Eedi platform and found that each tutoring conversation took roughly four to seven minutes of lesson time. Conversations moved slowest for students working with an expert tutor, who waited a median of 24 seconds for each tutor reply, compared with 16 seconds when receiving AI tutoring with question context and 19 seconds when receiving AI tutoring with personalised context. Correspondingly, over the 10 weeks of classroom activity, students completed a median of 99 study units in each of the AI tutoring conditions and 77 study units when working with an expert tutor, compared with 132 study units for students receiving standard remediation. Completing fewer study units meant that tutored students covered a narrower range of mathematical concepts in their work—a median of 57 distinct constructs in the AI tutoring conditions and 48 with an expert tutor, versus 72 with standard remediation. The effect of this coverage gap shows up clearly, independently of the effect of condition. That is, controlling for condition, each +1 SD increase in the number of study units a student completed corresponded to a +0.09 SD [+0.03, +0.16] gain on Star Maths at endline (see Appendix E).

A second factor emerged when we examined differences by student gender. Among students receiving

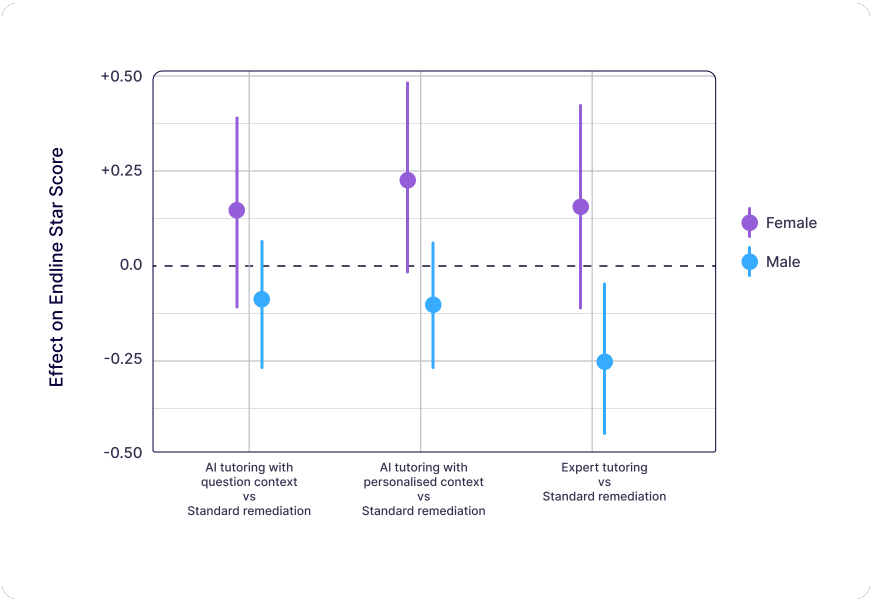

Figure 6 | Both expert and AI tutoring substantially narrowed the gender gap on Star Maths. Among students receiving standard remediation, boys outpaced girls from baseline to endline by +0.31 SD, whereas all three tutoring conditions reduced this gap to near zero. Across the tutoring conditions, girls improved relative to standard remediation, while boys did not yet show the same benefits. Points and error bars indicate posterior estimates and 95% credible intervals, respectively.

standard remediation, boys outpaced girls on Star Maths over the trial by +0.31 SD [+0.08, +0.54]. Both expert and AI tutoring substantially reduced this gender gap (Figure 6). AI tutoring with question context narrowed the relative advantage for boys to +0.09 SD [–0.08, +0.27], AI tutoring with personalised context decreased it to +0.02 SD [–0.17, +0.19], and expert tutoring reduced it to –0.05 SD [–0.28, +0.20]. This shift reflected directly opposed patterns for male and female students: across all three conditions, tutoring improved Star Maths scores for girls relative to standard remediation, whereas tutored boys tended to trail behind those receiving standard remediation (see Appendix E). Crucially, because the largest school participating in the trial was an all-boys school, boys made up two-thirds (66.1%) of the students in our sample. In short, tutoring helped girls on Star Maths, but its effects among boys pulled down the overall averages for the tutoring conditions.

Together, curriculum coverage and gender effects help make sense of the two puzzling patterns we observed. Accounting for completed study units and gender composition, the shortfall for expert tutoring dissipates, with an effect of –0.02 SD [–0.18, +0.16] relative to standard remediation. Meanwhile, students receiving AI tutoring with question context show an effect of +0.05 SD [−0.10, +0.20], and students receiving AI tutoring with personalised context an effect of +0.09 SD [−0.06, +0.23]. Under these adjustments, we estimate a 74.5% probability that AI tutoring with question context outperformed standard remediation, and an 87.9% probability that AI tutoring with personalised context improved upon standard remediation (see Appendix E).

Even if slower topic coverage and boys' lower gains help explain our Star Maths results, both factors raise the question of why tutoring would not face headwinds over the long term. For gender, we look to students' experiences of tutoring for potential insights (see Section 3.4). For breadth, platform metrics offer a direct clue. Students only receive support after making a mistake. Thus, as tutored students gradually improve their performance, they should enter fewer tutoring conversations. Over the course of the trial, accuracy on initial diagnostic questions indeed climbed steadily across each of the tutoring conditions, whereas it stagnated for students receiving standard remediation (see Figure 7). The slowdown in coverage may thus reflect an early cost—an investment in the future—rather than a lasting headwind. Taking a step back, the prior, independent two-year trial on the Eedi platform [8] offers a useful comparison. At its six-month mark, the trial found only small, uncertain effects on the standardised assessments that it used to measure broad attainment [30]. Clear overall learning gains emerged and stabilised around one to one-and-a-half years into the trial, affirming that pedagogical improvements may take time to accumulate.

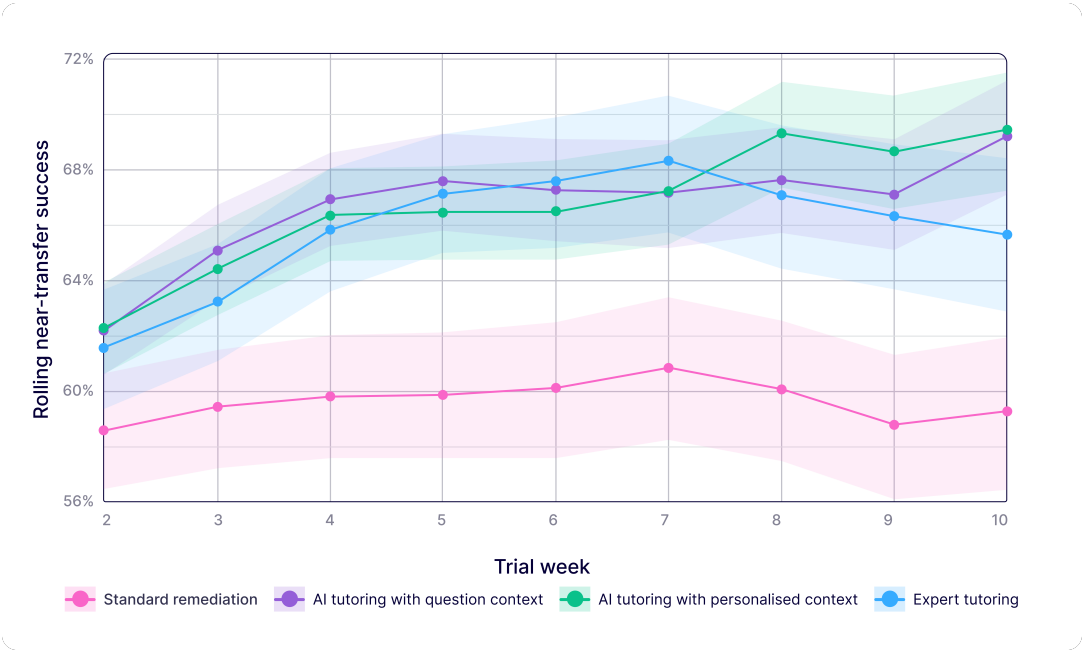

Figure 7 | Accuracy on initial diagnostic questions climbed steadily over the trial for tutored students, whereas it stagnated for students receiving standard remediation. Because the platform only triggers a tutoring chat after an error, rising first-attempt accuracy meant that tutored students needed fewer tutoring sessions as the trial progressed. Lines and shaded bands indicate posterior estimates and 95% credible intervals, respectively.

Returning to our pre-registered analyses, we next tested whether Star Maths gains varied with students' baseline attainment. Over 10 weeks of lessons, initial differences in Star Maths scores narrowed from baseline to endline for students receiving standard remediation or receiving AI tutoring, but persisted for students working with an expert tutor. Among students receiving standard remediation, on average, each +1 SD difference in baseline score narrowed to +0.74 SD [+0.62, +0.86] at endline. This initial gap narrowed similarly for students receiving AI tutoring: to +0.83 SD [+0.73, +0.92] with question context and +0.79 SD [+0.71, +0.89] with personalised context. In contrast, the gap did not narrow for students working with an expert tutor, with each +1 SD at baseline remaining +1.00 SD [+0.86, +1.08] at endline. We found no clear differences in the effects of each condition based on free-school-meal eligibility or school year group (see Appendix E).

3.4. Student experience of tutoring

What did students make of studying and working with tutors during this trial? During their tutoring sessions, students may have engaged deeply with the material, or leaned on the tutor to do the thinking for them [4, 31]. More broadly, sustained improvements may depend on deeper shifts in students' attitudes and habits as learners. We therefore analysed students' behaviour inside the tutoring conversations, as well as their responses to our baseline and endline surveys (**RQ5**).

We began by coding every tutoring transcript for genuine mathematical attempts, guessing at answer options, and requests to end the chat early (see Appendix F). When tutors attempted to initiate a chat, students sent a reply 76.0% to 77.0% of the time, depending on the condition. When they did reply, students made a genuine attempt at the maths in 85.4% to 86.0% of conversations across all three conditions. By comparison, students rarely tried to get an answer by guessing options without giving a reason (2.9% to 3.2% of AI tutoring conversations, and 6.5% with an expert tutor). Students did ask to leave early in roughly three out of 10 conversations: 29.5% to 29.6% in the AI tutoring conditions and 32.6% when working with an expert tutor. Over the course of the trial, however, these requests declined steadily from 38.3% of conversations in the first week to 23.9% in the final week. To follow up on the gender differences on Star Maths, we also compared these behaviours between boys and girls. Although overall exit rates varied across schools—and were highest at the all-boys school in our sample—boys and girls within the same school engaged and asked to leave at similar rates (see Appendix F). Student behaviour inside the tutoring conversations therefore does not seem to

account for the gender patterns we observed earlier.

We next turned to our baseline and endline surveys (n = 320 students completing both surveys) to check for deeper shifts in students' attitudes and habits as learners. These measured students' beliefs about learning from mistakes and adaptive reactions to mistakes [32], their mathematics self-efficacy [33], and their autonomous motivation to learn mathematics [34]. Although we hypothesised that guided support through mistakes would foster more constructive attitudes over time, students' responses on these measures changed very little over the 10-week trial. Across the survey sample as a whole, neither expert tutoring nor AI tutoring shifted students' attitudes toward mistakes, mathematics self-efficacy, or motivation relative to standard remediation (see Appendix F).

In exploratory analyses by student gender, however, we noticed overall patterns in students' mathematics self-efficacy and reactions to mistakes that parallel our findings on Star Maths. At baseline, boys scored higher than girls on both mathematics self-efficacy (3.66 and 3.03 out of 5, respectively) and adaptive reactions to mistakes (3.70 and 3.58 out of 5, respectively). Over the 10 weeks of the trial, both expert and AI tutoring narrowed these baseline gender gaps relative to standard remediation. This shift was clearest in students' reactions to making mistakes, where tutoring reduced the gap between boys and girls by 0.32 to 0.35 points (see Appendix F). Girls who received tutoring held steady or grew more comfortable working through mistakes over the trial, whereas tutored boys showed less comfort with mistakes than boys who received standard remediation. Having to pause and work through every mistake in a tutoring chat, rather than moving on after a brief hint, may therefore have created more initial friction for boys than for girls.

3.5. Personalised context in AI tutoring

Our next question centred on AI tutoring: did providing Gemini with personalised context on a student's learning history improve its tutoring (**RQ4**)? As our earlier results indicate, students in the two AI tutoring conditions performed very similarly across our proximate, intermediate, and broad learning measures (see Appendix E). On the student surveys, personalised context seemed to make little difference in students' attitudes: an effect of –0.01 points [–0.18, +0.17] on beliefs about learning from mistakes, +0.10 [–0.05, +0.27] on adaptive reactions to mistakes, +0.17 [–0.03, +0.37] on mathematics self-efficacy, and +0.32 [–0.49, +1.18] on autonomous motivation. Students also judged the two tutors similarly in effectiveness of understanding and adapting to their needs (2.73 and 2.84 out of 5, respectively; see Appendix F). Across both learning outcomes and student experience, then, our implementation of personalised context did not improve upon AI tutoring with question context alone.

Personalised context also seemed to make little difference to the tutoring conversations themselves. Supervising tutors approved Gemini's drafts without any changes at virtually the same rate when the model received personalised context as when it received question context alone (94.7% and 94.8% of reviewed drafts, respectively). When tutors did rewrite a draft, they intervened for the same reasons: adjusting tone and phrasing (33.3% and 31.8% of edits, respectively), refining the pedagogical approach (29.0% and 28.1%), or managing student engagement (21.9% and 23.8%; see Appendix D). During the chats, the two tutors withheld direct answers (98.7% and 98.8% of conversations) and asked follow-up questions to check understanding (91.5% and 91.2%) at nearly identical rates. Students also appeared to respond in consistent ways: they actively worked through the mathematics in 85.6% and 85.4% of their conversations, and asked to leave early in 29.5% and 29.6%. Assembling and processing a student's learning history did, however, slightly increase the time students waited for each tutor reply: a median of 19 seconds with personalised context, compared with 16 seconds with question context alone.

To make sense of this lack of differentiation, we traced the two steps of the personalisation process that take place before a chat begins: the summary of the student's learning history that the platform compiles for the prompt, and the initial teaching plan that Gemini writes to guide the session. Here, we found that the personalisation process did not play out quite as we had expected. First, in the learning-history summaries, a platform bug prevented notes from earlier tutoring chats from carrying over between sessions. As a result, Gemini received data on each student's proficiency, pace, and misconceptions across past study units, but no information on their previous tutoring conversations. Second, although this platform history shaped Gemini's initial teaching plans, Gemini only partially carried those plans through into the conversations (see Appendix G). When Gemini received personalised context rather than question context alone, its teaching plans cited the student's prior proficiency far more often (70.0% of conversations rather than 3.3%), as well as their pace

(67.8% rather than 0.8%). These plans affected Gemini's behaviour in some ways: for example, Gemini more often mentioned details from the student's history in its messages when provided with personalised context (18.4% of conversations, compared with 0.8% with question context alone). However, these changes did not seem to translate into pedagogically meaningful differences for students.

3.6. Tutor experience of AI supervision

Finally, we explored the perspectives and experiences that our cohort of expert tutors shared with us. Because each tutor rotated between shifts tutoring students directly and shifts supervising Gemini, their daily diaries allowed us to compare the two roles side by side (458 diaries, split between 271 supervisory shifts and 187 direct-tutoring shifts). In day-to-day practice, working with Gemini helped tutors support their students and bolstered their confidence in their teaching. Tutors reported that the messages drafted by Gemini were helpful in 81.9% of their supervisory diaries, that the suggestions saved them time (86.3%), and that they made composing messages less mentally taxing (77.9%). Comparing the supervisory and direct-tutoring shifts, tutors also reported higher confidence in their tutoring on supervisory shifts, with an average rating of 4.47 [4.40, 4.54] out of 5 relative to 4.34 [4.26, 4.42] after a direct-tutoring shift (posterior probability of increase: 99.4%). In contrast, tutors did not find it much easier to context-switch between concurrent student conversations, with an average of 4.48 [4.41, 4.55] relative to 4.46 [4.38, 4.54] (posterior probability of increase: 65.5%). In fact, overseeing the AI tutor appeared to carry its own demands, as tutors reported a decreased ability to manage the workload on supervisory shifts, at 4.33 [4.26, 4.41] relative to 4.43 [4.34, 4.52] (posterior probability of decrease: 94.7%).

As part of their daily diaries, tutors also completed the Positive and Negative Affect Schedule [35], giving us insight into their emotions throughout the trial. Supervising Gemini appeared to dampen the positive emotion tutors felt in their work. Positive affect averaged 3.59 [3.52, 3.65] out of 5 after supervisory shifts, compared with 3.76 [3.68, 3.84] after direct-tutoring shifts (posterior probability of decrease: >99.9%). The drop was clearest in feelings of excitement, pride, strength, and determination (see Appendix H). Negative affect remained very low in both roles, with a slight rise after supervisory shifts: 1.17 [1.14, 1.20] out of 5, compared with 1.12 [1.08, 1.15] after direct-tutoring shifts (posterior probability of increase: 98.0%), concentrated in feelings of guilt and irritability.

Beyond these daily check-ins, we surveyed tutors at baseline and endline (n = 40 tutors completing both surveys) to investigate whether any of their broader attitudes and beliefs shifted over the course of the trial. On the whole, tutors entered the trial with fairly favourable views of AI, and largely maintained those views at endline. Tutors expressed strong trust in the AI tutor's mathematical accuracy, with an average rating of 4.05 [3.84, 4.25] out of 5 at baseline and 4.14 [3.92, 4.35] at endline (posterior probability of increase: 79.2%). Tutors also agreed with the idea that AI tools could help support their teaching practice, with an average of 4.02 [3.76, 4.29] at baseline and 3.99 [3.73, 4.26] at endline (posterior probability of increase: 42.1%). By contrast, tutors communicated concerns about potential drawbacks for their role—with some evidence of this ambivalence increasing over the trial. Specifically, tutors shared wariness that relying on AI could weaken their own tutoring skills, with an average of 2.96 [2.61, 3.31] at baseline and 3.07 [2.74, 3.43] at endline (posterior probability of increase: 73.6%). Tutors also voiced mild concern that the AI tutor was doing work that should be done by a human, at 2.59 [2.36, 2.83] at baseline and 2.71 [2.48, 2.94] at endline (posterior probability of increase: 81.3%; see Appendix H).

Following the trial (n = 8 tutors participating in endline interviews), tutors described a positive overall experience working with Gemini. In six of the interviews, the tutors placed particular emphasis on the ability to support more students at once while maintaining the quality of their teaching. Reviewing the AI tutor's drafts strengthened tutors' own practice as well: all but one of the tutors mentioned drawing inspiration from its tutoring explanations when later working with students directly. In fact, in five of the interviews, tutors specifically mentioned adopting Gemini's habit of asking a follow-up question to check student understanding in their own tutoring sessions. At the same time, seven of the eight tutors noted that tracking several live chats at once remained challenging. Most also spoke of the personal fulfilment they drew from teaching students one-to-one. Ultimately, the tutors underscored that human teachers remain essential alongside AI, particularly for diagnosing students' misconceptions. Appendix H presents the complete codebook and the full set of themes from our analysis.

4. Discussion

Last year, our 2025 exploratory RCT gave us an encouraging first look at AI tutoring across nine mathematics classrooms in the UK [1]. Our new, pre-registered RCT expanded our evaluations to 58 classrooms and $N = 1179$ students, validating that initial promise at scale. To build our tutor, we integrated Gemini into the Eedi mathematics platform, where it drafted tutoring messages for students under the direct supervision of $N = 58$ expert educators. Throughout the trial, the AI tutoring consistently met the teaching standards of the supervising educators: the AI tutor reliably followed our pedagogical guidelines, rarely made mathematical mistakes, and maintained an appropriate tone in its messages. For students, this support translated into material improvements across our proximate and intermediate measures of learning. During their weekly lessons on the platform, students tutored by Gemini recovered from their mistakes and tackled new topics more successfully than peers who received the platform's standard remediation. These proximate gains persisted over subsequent weeks: students tutored by Gemini also demonstrated clear gains on an intermediate midtrial quiz. Notably, on each of these measures, Gemini tutoring performed just as effectively as tutoring by expert educators.

In contrast, neither Gemini tutoring nor tutoring by expert educators resulted in clear improvements over standard remediation on our broad measure of learning, the Renaissance Star Maths assessment. While we knew that educational interventions tend to show smaller effects on broad standardised tests than on curriculum-aligned assessments [36, 37], these results still surprised us. We hypothesised that the benefits of AI tutoring would fall short of expert tutoring, not for the latter to underperform both AI support and standard remediation. Our investigation of this puzzle revealed a sharp trade-off between depth and breadth during classwork on the Eedi platform (see also [29]). Working through a mistake with a tutor takes several minutes longer than reading a short hint, so tutored students covered fewer topics over the 10 weeks—leaving them with less exposure to the wide curriculum tested by Star Maths. However, over the trial, tutored students grew steadily more accurate when tackling new topics for the first time. In contrast, the students receiving standard remediation plateaued. The slower initial pace of tutoring may act as an upfront investment in deeper understanding that takes time to pay dividends.

Both expert and AI tutoring narrowed baseline gender gaps in students' mathematics attainment. One-to-one tutoring helped girls make progress on their studies, but seemed to create friction for some boys. This may have resulted from boys entering the trial more confident in their maths abilities: for self-assured students, being diverted into a mandatory chat after every mistake may have registered as an unwelcome signal of failure [32, 38]. Indeed, we observed an initial gender gap in students' adaptive reactions to making mistakes, which later constricted—specifically in the three tutoring conditions. Conversational tutoring may therefore not be a panacea for every error. What are the conditions that define when a student would genuinely benefit from conversational tutoring, and when a brief hint would suffice instead?

One particularly notable difference between AI and expert tutoring was the use of follow-up questions. At the end of its chat sessions, Gemini asked follow-up questions to verify student understanding far more frequently than expert tutors did. Indeed, in interviews after the trial ended, multiple supervising tutors mentioned that overseeing the AI tutoring sessions prompted them to adopt the habit of asking these questions into their own practice. These extra checks on understanding may have proved particularly helpful for lower-attaining students—and may partly account for the narrowing of the initial attainment gap, which otherwise persisted under expert tutoring. At the same time, probing further after correct answers added friction for students who did not want to be in a chat in the first place. Supervising tutors frequently stepped in at the close of conversations to let students move on. Again, navigating this tension—identifying when to press for deeper understanding and when to let a student move on—seems like a crucial research challenge for AI tutoring.

One distinctive feature of our trial design was its use of an active, evidence-based control condition. Standard remediation on the Eedi platform uses diagnostic questions to identify the likely misconceptions behind students' mistakes and deliver educator-designed hints, fluency exercises, and videos tailored to those misconceptions. Prior classroom RCTs show that this approach is an effective learning intervention in its own right [39, 8]. Over two years, students assigned to use the platform achieved learning gains equivalent to roughly four additional months of progress relative to regular classroom instruction. Testing AI tutoring against such a challenging benchmark holds it to a practical standard for schools: matching or improving upon proven interventions.

Our approach to personalisation did not appear to make a substantial difference in the AI tutoring process.

This may partly reflect the strength of the platform's diagnostic questions. Once a question identifies a candidate misconception behind a mistake, a summary of a student's past work may add little extra value. Another possibility, of course, is that our implementation fell short of encoding the important aspects of personalisation. As a result, we interpret our results with caution. Future studies should continue exploring ways to adapt AI tutoring to each student's strengths, struggles, and progress over time [18].

For the expert educators participating in our trial, supervising Gemini carried a clear trade-off. Because virtually all of Gemini's messages met high pedagogical standards, supervising tutors rarely needed to rewrite its drafts. This reliability saved tutors time, eased the mental effort of composing replies, and boosted their confidence in their tutoring. Yet monitoring several live AI chats at once carried its own cognitive demands [40, 41], and tutors reported greater difficulty managing their overall workload than when tutoring students directly. The shift from crafting explanations to reviewing ready-made drafts also dampened tutors' daily feelings of pride and excitement, consistent with broader research showing that people find active collaboration with AI more fulfilling than passive review [42, 43]. These findings point to the value of keeping teachers at the centre: drawing on the complementary strengths of human educators and AI tutoring, while protecting the personal engagement with students that makes teaching rewarding.

Overall, this trial offers further evidence that pedagogically aligned AI can safely and effectively support students in real classrooms. At the same time, our attempt to trace student progress from immediate mistakes and recovery to broad learning points to a key tension for those building the burgeoning evidence base for AI in education [3]. Students develop proficiency and mastery over the course of months and years. Modern AI capabilities, in comparison, often advance much more quickly. Bridging these two timescales will require new research innovations—agile studies to evaluate the rapid evolution of AI models, as well as longer studies to track students' cumulative attainment.

5. Methods

We submitted our protocol for independent ethical review, and received a favourable opinion from the Human Behavioural Research Ethics Committee at Google DeepMind (#25 003).

Participants We recruited $N = 1179$ students from 10 state-funded secondary schools in Eedi's testbed programme. These schools included six in England, two in Northern Ireland, one in Wales, and one in Scotland. All of these schools were already active Eedi users: that is, even preceding the trial, they incorporated the Eedi platform as part of their mathematics instruction. Across these schools, we drew our sample from 58 mathematics classes (see Appendix B for school and sample details). Participating students came from Years 8 to 10 in England and Wales, along with the equivalent year groups in Northern Ireland (Years 9 to 11) and Scotland (S2 to S4). Across all schools, these year groups corresponded to ages 12 to 15. During the trial, these classes continued to use the platform in their regular mathematics lessons for roughly 30 to 50 minutes per week. Following discussions with the participating schools, a legitimate interests assessment, and consultation with our ethics committee, we adopted an opt-out process for the trial. Parents and students could opt out of the research at any time. Before the trial began, the schools shared study information with parents and students, and sent several follow-up reminders about the upcoming study and their option to opt out. As part of this process, we explained to parents and students that some tutoring during the trial would involve AI support. Finally, we recruited $N = 58$ expert tutors—all experienced mathematics educators—for the trial. Each tutor provided informed consent to participate in the research.

Platform During the trial, the Eedi platform organised student activity into short study units focused on specific mathematics topics. Each study unit included two diagnostic multiple-choice questions. Each diagnostic question presented four response options. One option was the correct answer, and the other three were distractors mapped to specific mathematical misconceptions. When a student answered a unit's initial diagnostic question correctly, the platform advanced them directly to the next study unit. When a student answered the initial diagnostic question incorrectly, the platform triggered the support for their assigned condition. Unlike in our 2025 exploratory trial, the platform did not prompt students to retry the identical question after completing support. Instead, the platform moved students directly to the unit's second diagnostic question. This second question assessed the same topic and misconceptions using different numbers or details. Once a student

answered this second question, the platform completed the study unit and advanced them to the next unit. Beyond the support triggered automatically after an error, the platform included an "Ask for help" button across all four conditions. The platform specifically removed this button when presenting the second diagnostic question in a study unit. Otherwise, students could click the button at any point to request a live chat with a tutor (see Appendix B).

Model To support the two AI tutoring conditions, the Eedi platform called Gemini via API. As new models launched during the trial, we updated the API routing from Gemini 3.1 Pro (from the beginning of the trial until 21 April) to Gemini 3 Flash (from 21 April until 21 May) and Gemini 3.5 Flash (from 21 May until the end of the trial). At the start of each conversation, the platform prompted Gemini to write a short teaching plan to outline an instructional strategy to attempt to address the student's likely misconception. To draft each tutoring message in the conversation, the platform prompted Gemini with the core pedagogical prompt, the session's teaching plan, the conversation transcript so far, and either the "question context" or the question context and "personalised context".

In both AI tutoring conditions, the platform supplied the question context, or information on the study unit the student was working on. Specifically, the platform provided the diagnostic question, the four answer options, the correct solution, an educator-written explanation of the inferred misconception, and the student's position and activity in the unit so far. In the personalised condition, the platform additionally supplied the personalised context, or a profile of the student's prior activity on Eedi. This profile included the student's school year group, their recent study-unit scores and effort, their predicted proficiency, and their past errors on prerequisite topics. We intended this profile to incorporate summaries of the student's earlier tutoring chats as well, but a technical bug prevented the platform from carrying those chat summaries over between sessions (see Appendix C and Appendix G).

Procedure We took two preparatory steps before launching the actual trial. First, we ran a brief, in-classroom test of our trial protocols beginning on 23 March. Second, we pre-registered our study design and analysis plan with the AEA RCT Registry (`AEARCTR-0018079` [27]). We then launched the trial on 13 April. Excluding a week-long school holiday and the endline testing period, the trial covered 10 weeks of classroom instruction, ending on 7 July. While most schools completed the full 10 weeks, instruction ranged from seven to 10 weeks due to differing term dates and endline testing schedules. To ensure that differences between classrooms would not confound our comparisons between conditions, we randomised at the student level. That is, we randomly assigned students within each classroom to one of the four conditions for the full duration of the trial. We assigned a larger share of students to the two AI tutoring conditions to improve our precision when analysing the effects of AI tutoring, targeting an overall ratio of 3:5:5:3 for the relative size of our four conditions. In total, we assigned 220 students to standard remediation, 368 to AI tutoring with question context, 371 to AI tutoring with personalised context, and 220 to expert tutoring (see Appendix B).

To support the tutoring conditions during school hours, we scheduled tutors to rotate across three types of daily shifts. Depending on their shift, tutors answered student-initiated help requests from standard remediation, supervised Gemini across the two AI tutoring conditions, or wrote messages from scratch in expert tutoring. Across all shifts, tutors managed up to six conversations at a time through the Eedi tutoring dashboard. On supervisory shifts, the dashboard displayed the diagnostic question, the session's teaching plan, and Gemini's draft message. The supervising tutor could then approve the draft unchanged, prompt Gemini to regenerate it, or rewrite it before delivery. To keep tutors blind to our primary hypotheses, we gave the shift types neutral labels on the dashboard (that is, we made no explicit references to personalisation). For students in the two AI tutoring conditions, the chat window included a clear indicator showing that their tutoring involved AI support.

Throughout the trial, the Eedi platform recorded the diagnostic questions that students attempted during their lessons. Specifically, for each attempt, the platform recorded a pseudonymised, trial-specific identifier for the student, the question itself, its position in the study unit, the timestamp, the student's answer, and whether the answer was correct. Across all tutoring conversations, the platform recorded the messages exchanged between students and tutors, along with the message timestamps. In the AI tutoring conditions, the platform also recorded every draft generated by Gemini and the supervising tutor's decision (approve, regenerate, or rewrite). Outside the platform, teachers administered the eight-question curriculum quiz to their classes partway through the trial. Renaissance independently administered and scored the Star Maths assessment at

baseline (February to March) and at endline (June to July). Students completed the assessments during their regular mathematics lessons. Because student attendance varied across lessons and testing days, the number of students with complete data differed across our three levels of learning measures (see Appendix B).

To better understand student and tutor perspectives and experiences, we incorporated student and tutor surveys, daily tutor diaries, and tutor interviews into our trial protocol. For students, we administered surveys at baseline and endline. These surveys included instruments to measure students' beliefs about learning from mistakes and positive reactions to making mistakes [32], confidence in their mathematics ability [33], and autonomous motivation to learn the subject [34] (see Appendix F). For tutors, we also administered surveys at baseline and endline. These surveys included questions to measure tutors' attitudes and beliefs about using AI in tutoring. Throughout the trial, tutors completed a brief daily diary after each shift. In these diaries, tutors rated their confidence in their tutoring, the ease of switching between concurrent conversations, their ability to manage their workload, and the positive and negative emotions they experienced during the shift [35]. After supervisory shifts, the diary also prompted tutors to rate the helpfulness of the messages that Gemini drafted that day. Finally, at the end of the trial, we conducted a series of in-depth, semi-structured interviews with eight tutors (see Appendix H). We pseudonymised all study data after collection so that the research dataset contained no direct personal identifiers.

Analysis We evaluated student learning across our three levels of outcomes: proximate performance (immediate mistake correction on the second diagnostic question of a study unit and near transfer to the initial diagnostic question of the next study unit), intermediate consolidation (scores on the eight-question midtrial quiz), and broad attainment (endline Star Maths scores). Across all learning outcomes, we estimated treatment effects using Bayesian hierarchical regressions. We included students' baseline Star Maths performance as a covariate in all primary models to account for pre-existing differences between students. To account for clustering in the data, the models also included random intercepts for classrooms. For the proximate platform outcomes, given the repeated measurements for each student, the models additionally included random intercepts for individual students. In *Results*, the accuracy rates reported for the proximate and intermediate outcomes represent baseline-adjusted posterior predictive margins from these regressions. For Star Maths, we standardised scores across the sample and report baseline-adjusted differences between conditions in standard deviation units (see Appendix E).

We used a similar Bayesian approach to analyse our survey and diary data. For the student surveys, we estimated treatment effects on endline scores using Bayesian linear regressions that adjusted for each student's baseline score on the same measure (see Appendix F). For the tutor surveys and daily diaries, we used Bayesian hierarchical regressions to compare tutors' responses between baseline and endline and between supervisory and direct-tutoring shifts. Because each tutor provided multiple responses over the trial, these models included random intercepts for individual tutors (see Appendix H).

To evaluate the safety and pedagogical quality of the AI tutor's drafts, we combined human review with automated coding. For drafts that supervising tutors edited or rewrote, we quantified the size of each revision by computing the character-level Levenshtein distance and the edit ratio (the Levenshtein distance divided by the length of the initial draft). We next inferred the functional purpose of each revision. We enlisted several expert tutors from the trial to conduct three rounds of review, in which they proposed labels for pairs of Gemini drafts and tutor revisions. This review process produced an eight-category codebook. We then used Gemini 3.7 Flash to apply this codebook across the entire corpus of revised drafts. We also audited the tutoring drafts and conversations for tone appropriateness and mathematical accuracy. Specifically, we applied Gemini 3.7 Flash to flag potential mathematical errors and tone or safety issues across every AI tutor draft. In addition, we used Gemini 3.8 Flash to flag potential contextual mathematical errors across all expert tutoring conversations and all AI tutoring conversations that ran without supervisor revisions. Members of the research team then manually reviewed and adjudicated every flagged message (see Appendix D).

We conducted three additional analyses to examine tutor pedagogy, student behaviour, and session planning. First, we used Gemini 3.8 Flash to evaluate adherence to our pedagogical guidelines across all tutoring conversations in which the student replied (see Appendix D). Second, we leveraged Gemini 3.7 Flash to evaluate student behaviour in tutoring conversations across 17 binary indicators. Before coding, we redacted the assigned condition and any information about the tutor from the transcripts. To validate the classifications, members of the research team manually coded a subset of conversations (see Appendix F). Lastly, to examine

the effect of personalised context on session planning, we analysed the teaching plans that Gemini generated at the start of each conversation. Specifically, we used a rule-based keyword search to identify references in these plans to the student's prior proficiency or pace (see Appendix G).

Finally, we analysed the eight tutor interview transcripts using an iterative thematic analysis [44]. Members of the research team first independently reviewed an initial subset of transcripts to generate candidate codes. The team then consolidated these codes into a structured codebook spanning 57 codes and nine overarching themes. Working in small groups, researchers then independently coded each of the eight transcripts using this codebook and met to reconcile any differences (see Appendix H).

References

[1] LearnLM Team, Google and Eedi. AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. *arXiv preprint arXiv:2512.23633*, 2025. URL https://arxiv.org/abs/2512.23633.

[2] LearnLM Team. LearnLM: Improving Gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024. URL https://arxiv.org/abs/2412.16429.

[3] Lily Fesler, J. Pablo Martinez Claeys, Chris Agnew, and Susanna Loeb. The evidence base on AI in K–12: A 2026 review. Technical report, AI Hub for Education, SCALE Initiative, Stanford University, 2026. URL https://scale.stanford.edu/research-in-action/understanding-evidence-base-ai-k12-education.

[4] 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. doi: 10.1073/pnas.2422633122.

[5] 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. doi: 10.1038/s41598-025-97652-6.

[6] Eamon Worden, Luca Dang, Wen-Chiang Ivan Lim, Sarah Miller, Jiayi Zhang, Aaron Haim, Adam Sales, Ashish Gurung, and Neil T. Heffernan. A large scale randomized control trial showing LLM generated feedback helps low-knowledge middle school math students with short-term learning. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale*, pages 361–372, 2026. doi: 10.1145/3774398.3811594.

[7] Philip Oreopoulos, Michael Liut, Alp Sungu, and Nina Low. Making AI tutoring productive: Evidence from a mastery-based math practice experiment. EdWorkingPaper 26-1552, Annenberg Institute at Brown University, 2026. URL https://doi.org/10.26300/01qv-6c22.

[8] Emma S. Dobson, Rahil Khowaja, Wayne Harrison, Germaine Uwimpuhwe, and Steve Higgins. The impact of sustained use of Eedi on KS3 mathematics attainment: Evidence from a two-year cluster randomised controlled trial (Years 7–8). Technical report, WhatWorked Education, Durham, UK, 2026.

[9] E. Caroline Wylie and Dylan Wiliam. Diagnostic questions: Is there value in just one? In *Annual Meeting of the American Educational Research Association (AERA)*, 2006.

[10] Philip M. Sadler. Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35(3):265–296, 1998. doi: 10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P.

[11] David F. Treagust. Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, 10(2):159–169, 1988. doi: 10.1080/0950069880100204.

[12] Derek C. Briggs, Alicia C. Alonzo, Cheryl Schwab, and Mark Wilson. Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11(1):33–63, 2006. doi: 10.1207/s15326977ea1101_2.

[13] John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray Pelletier. Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2):167–207, 1995. doi: 10.1207/s15327809jls0402_2.

[14] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011. doi: 10.1080/00461520.2011.611369.

[15] Andre Nickow, Philip Oreopoulos, and Vincent Quan. The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, 61(1):74–107, 2024. doi: 10.3102/00028312231208687.

[16] Michela Carlana and Eliana La Ferrara. Apart but connected: Online tutoring, cognitive outcomes, and soft skills. *American Economic Review*, 115(10):3487–3513, 2025. doi: 10.1257/aer.20240401.

[17] Lucas Gortazar, Claudia Hupkau, and Antonio Roldán-Monés. Online tutoring works: Experimental evidence from a program with vulnerable children. *Journal of Public Economics*, 232:105082, 2024. doi: 10.1016/j.jpubeco.2024.105082.

[18] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994. doi: 10.1007/BF01099821.

[19] LearnLM Team. Evaluating Gemini in an arena for learning. *arXiv preprint arXiv:2505.24477*, 2025. URL https://arxiv.org/abs/2505.24477.

[20] Renaissance Learning. Star Assessments for Maths: Technical manual. Technical report, Renaissance Learning, London, UK, 2019.

[21] Malcolm Swan. Dealing with misconceptions in mathematics. In Peter Gates, editor, *Issues in Mathematics Teaching*, pages 147–165. RoutledgeFalmer, London, 2001.

[22] Julie L. Booth, Karin E. Lange, Kenneth R. Koedinger, and Kristie J. Newton. Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction*, 25:24–34, 2013. doi: 10.1016/j.learninstruc.2012.11.002.

[23] James M. Royer. Theories of the transfer of learning. *Educational Psychologist*, 14(1):53–69, 1979. doi: 10.1080/00461527909529207.

[24] David N. Perkins and Gavriel Salomon. Transfer of learning. In Torsten Husén and T. Neville Postlethwaite, editors, *The International Encyclopedia of Education*, volume 11, pages 6452–6457. Pergamon Press, Oxford, 2nd edition, 1994.

[25] Nicholas C. Soderstrom and Robert A. Bjork. Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2):176–199, 2015. doi: 10.1177/1745691615569000.

[26] Doug Rohrer and Kelli Taylor. The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology*, 20(9):1209–1224, 2006. doi: 10.1002/acp.1266.

[27] Vasco Brazão. Testing the efficacy of AI tutoring in secondary mathematics: A scaled RCT in UK classrooms. AEA RCT Registry, AEARCTR-0018079, April 2026. URL https://www.socialscienceregistry.org/trials/18079.

[28] Education Policy Institute and Renaissance Learning. Understanding progress in the 2020 to 2021 academic year: Complete findings from the autumn term. Research Report DFE-RR1133, Department for Education, 2021. URL https://www.gov.uk/government/publications/pupils-progress-in-the-2020-to-2022-academic-years.

[29] Shayan Doroudi, Vincent Aleven, and Emma Brunskill. Where's the reward? A review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education*, 29(4):568–620, 2019. doi: 10.1007/s40593-019-00187-x.

[30] Northwest Evaluation Association. MAP Growth technical report. Technical report, Northwest Evaluation Association (NWEA), Portland, OR, 2019. URL https://www.nwea.org/uploads/2021/11/MAP-Growth-Technical-Report-2019_NWEA.pdf.

[31] Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker, and Rachit Dubey. AI assistance reduces persistence and hurts independent performance. *arXiv preprint arXiv:2604.04721*, 2026. URL https://arxiv.org/abs/2604.04721.

[32] Maria Tulis, Gabriele Steuer, and Markus Dresel. Positive beliefs about errors as an important element of adaptive individual dealing with errors during academic learning. *Educational Psychology*, 38(2):139–158, 2018. doi: 10.1080/01443410.2017.1384536.

[33] Paul R. Pintrich, David A. F. Smith, Teresa Garcia, and Wilbert J. McKeachie. Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, 53(3):801–813, 1993. doi: 10.1177/0013164493053003024.

[34] Richard M. Ryan and James P. Connell. Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57(5):749–761, 1989. doi: 10.1037/0022-3514.57.5.749.

[35] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6): 1063–1070, 1988. doi: 10.1037/0022-3514.54.6.1063.

[36] James A. Kulik and J. D. Fletcher. Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1):42–78, 2016. doi: 10.3102/0034654315581420.

[37] Matthew A. Kraft, Beth E. Schueler, and Grace T. Falken. What impacts should we expect from tutoring at scale? Exploring meta-analytic generalizability. *Review of Educational Research*, 2026. doi: 10.3102/00346543261446660.

[38] Allison M. Ryan, Paul R. Pintrich, and Carol Midgley. Avoiding seeking help in the classroom: Who and why? *Educational Psychology Review*, 13(2):93–114, 2001. doi: 10.1023/A:1009013420053.

[39] 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, August 2024. URL https://www.interventions.whatworked.education.

[40] Joel S. Warm, Raja Parasuraman, and Gerald Matthews. Vigilance requires hard mental work and is stressful. *Human Factors*, 50(3):433–441, 2008. doi: 10.1518/001872008X312152.

[41] Raja Parasuraman and Dietrich H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 2010. doi: 10.1177/0018720810376055.

[42] Suqing Wu, Yukun Liu, Mengqi Ruan, Siyu Chen, and Xiao-Yun Xie. Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. *Scientific Reports*, 15(1):15105, 2025.

[43] Elena Hayoung Lee, Yidan Yin, Nan Jia, and Cheryl J. Wakslak. Relying on AI at work reduces self-efficacy, ownership, and meaning while active collaboration mitigates the effects. *Scientific Reports*, 16(1):13583, 2026. doi: 10.1038/s41598-026-42312-6.

[44] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.

Contributions and Acknowledgements

The following individuals contributed to the work described in this report. Within each group, we list names alphabetically rather than by rank of contribution.

Core Contributors

On the Google team, the following individuals made sustained contributions across the project:

Albert Wang, Irina Jurenka, Jonathan Caton, and Kevin R. McKee.

On the Eedi team, the following individuals made sustained contributions across the project:

Ariella Shulman, Bibi Groot, James Stalley-Moores, and Vasco Brazão.

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

Contributors

On the Google team, the following individuals contributed to specific parts of the trial:

Elizabeth Lapo, Fatima Mora Garcia, Katie Stasaski, Khyati Jain, Markus Kunesch, Miruna Pîslar, Rishabh Trivedi, and Yilmazcan Ozyurt.

On the Eedi team, the following individuals contributed to specific parts of the trial:

Anna Kenolty, Ben Caulfield, Beth Lilley-Draper, Chelsea Burdett, Claire Willis, Digory Smith, Ewan Stamford, Harriet Walters, Kelvin Pang, Lucy Dalton, Mike Beech, Nadia McLaren, Pauline Malubay, Rachel Kidson, Sam Wheeler, and Tisha Verma.

Acknowledgements

This work represents a close collaboration between Google and Eedi.

For Eedi: We thank Craig Barton, Gemma Bazany-Barber, Prarthana Bhattacharyya, Piri Boland, Niamh Bushnell, Samantha Headford, Iris Hulls, Lorna Kavanagh, Nigel Kendall, Lauren Leibman, Josh Mitton, George Mu, Anjali Nambiar, Rebecca Sudan, Rich Wells, Simon Woodhead, and Matthew Zent for their contributions to the trial and report; the Eedi tutors who supervised and delivered tutoring throughout the trial; Ryan Baker for advice on the simulation and trial design; and the students, teachers and heads of mathematics at the ten participating schools. Participating schools included Bullers Wood School for Boys, The Cheadle Academy, Mark Rutherford School, and St. Luke's High School (Barrhead, Glasgow). We thank Lisa-Mari Krige, Elizabeth Ball, Rebecca Kirkman, James Doughty and Graham Baird for their support of the trial in their schools.

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 Antonia Mould, Chloe Conacher, Diana Carreno, Emma Brunskill, Kaiz Alarakyia, Laura Breen, Liam McCafferty, Miriam Schneider, Sabrene Ben Dekhil, Samantha Heyman, and Shanice Onike.

Furthermore, we would like to thank the Gemini team and the Responsible Development and Innovation team at Google DeepMind. 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, Chris Phillips, Lila Ibrahim, Shakir Mohamed, and Zoubin Ghahramani.

Appendix

We plan to release a detailed appendix to accompany this report, as we did for the report describing our exploratory RCT [1]. Given the number of analyses we conducted for this trial, it will take a bit more time to compile and finalise the appendix. In the meantime, if you have an urgent question about our analysis or any other detail in the paper, please reach out to us at learnlm-tech-report@google.com and hello@eedi.com. Otherwise, we look forward to sharing the full appendix with you soon.