STUDENTBENCH: AI AND HUMAN TUTORING YIELD EQUIVALENT GRE LEARNING GAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI offers an unprecedented opportunity to augment human capabilities by teaching us new knowledge and skills. However, progress at the frontier has focused primarily on advancing model capabilities. We introduce StudentBench, a suite of LLM teaching evaluations and a public platform that enables large-scale data collection with over 175,000 student–AI messages to study whether LLMs produce learning gains equivalent to human tutoring. Using StudentBench, we measured learning gains on Quantitative and Verbal GRE questions across 2,383 human participants receiving AI tutoring, human tutoring, or no tutoring. We establish that pooled AI tutoring is statistically equivalent to expert human tutoring for combined GRE learning gains (p=.015). In a second study, expert human tutors compared LLM-generated lesson plans and practice problems through 2,028 pairwise rubric evaluations. The two studies clearly separate AI tutors across (1) lesson planning, (2) practice-problem creation, (3) conversational pedagogy, (4) cost, and (5) engagement. One model achieved learning gains equivalent to human tutoring (p=.044) at 918× lower cost per percentage point of learning gain. For Quantitative GRE sessions, faster AI replies correlated with more student messages, more messages with more correct practice, and more correct practice with larger learning gains.

1 THE PURSUIT OF HUMAN-LEVEL TUTORING

A singularity in intelligence can occur when an AI technology recursively self-improves, such that there is an irreversible split in intellectual ability between those who embrace the technology and those who do not (Good, 1965). While work in artificial intelligence has focused on the historical moment of recursive self-improvement for AI itself (Vinge, 1993; Zhang et al., 2026a;b), we examine a parallel singularity for human intelligence.

Humans augmented by teaching technologies may recursively self-improve faster than those who do not use them. We use the term *recursive human self-improvement* (RHSI) for this process: better teaching technologies help people learn, and greater human capabilities can support better use of those technologies (Engelbart, 1962). Establishing RHSI requires three conditions: first, the technology improves human capabilities; second, an improved human is better at using the technology; and third, better use leads to consistent further gains in human capabilities. If these conditions are met, humans using the technology meet the requirements to recursively self-improve.

Here, we investigate the first condition for RHSI, using human tutoring as a benchmark to test whether current LLMs can produce equivalent learning gains given one hour of teaching time. We make an experimental design choice to use minimal software scaffolding and two low-guidance prompts to prioritize measuring LLM teaching ability over our prompting or software design ability. We designed the study to compare human and AI tutors against the same no-tutoring control, randomly counterbalancing (swapping) the pre-test and post-test assessment forms so that differences in test difficulty would not be mistaken for learning gains (Section 2; Appendix A).

We focus our study on the Graduate Record Examination (GRE) for its widespread adoption (Educational Testing Service, 2026), coverage of mathematical skills also assessed on the SAT (Educational Testing Service, n.d.; College Board, 2024), and role as a gatekeeping mechanism between students and life opportunities such as graduate school, business school, and careers that further education can open (Educational Testing Service, 2026; Miller et al., 2019). Our GRE assessments cover seven

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

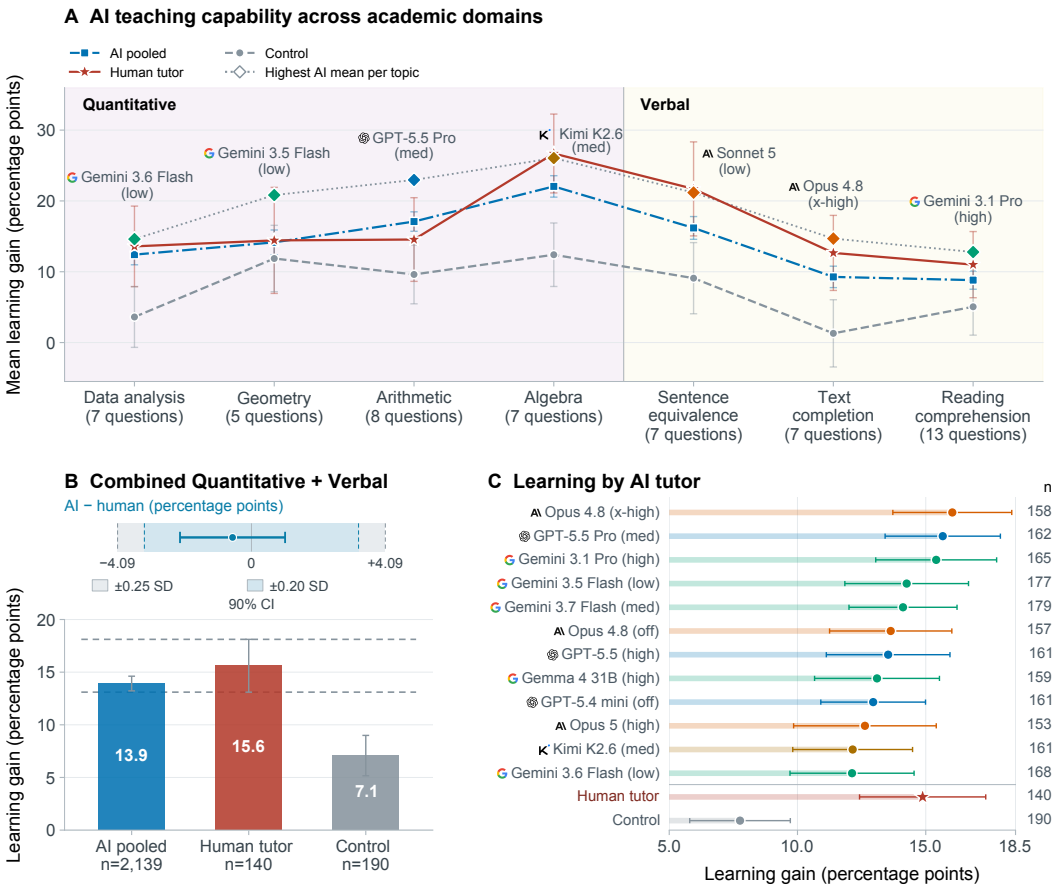

Figure 1: **Learning across GRE domains and tutors. A,** Mean pre- to post-test score change within each GRE domain under different tutoring conditions: human tutoring, AI tutoring (results pooled across all AI tutors), or no-tutoring control. Error bars depict 95% Student-t confidence intervals. Question counts refer to each assessment (pre-test, post-test). **B,** Learning gains (post-test − pre-test score) pooled across Quantitative and Verbal, with 95% intervals; n denotes the number of sessions under each tutoring condition. Dashed guides indicate the confidence interval for the gain from human tutoring. The inset shows the adjusted AI − human difference and 90% confidence interval within ± 0.25 pooled-SD equivalence bounds (p=.015); the interval also lies within the stricter ± 0.20-SD bounds (p=.023). **C,** Learning gains for the 12 AI tutors tested in both sections, adjusted for pre-test score and section, with 95% HC3 intervals. See Appendix C for learning gains (Table C.1) and Appendix C.2 for AI tutor leaderboards for all seven domains.

domains across Quantitative and Verbal reasoning, enabling exploration of the frontier of current LLM teaching capabilities across a variety of topics.

Earlier intelligent tutoring systems have produced learning outcomes comparable to human tutoring (VanLehn, 2011; Ma et al., 2014), and recent studies evaluate structured AI tutors and AI assistance to human tutors (Kestin et al., 2025; Wang et al., 2025; LearnLM Team, Google and Eedi, 2025). StudentBench tests autonomous, general-purpose LLMs with minimal software scaffolding against live human tutors on unaided GRE assessments; see Section 7 for further related work.

We establish the first condition for RHSI through one-hour AI tutoring sessions on GRE questions. Across 2,383 participants, with 13 AI tutors in each GRE section, we show that pooled AI tutoring and human tutoring produced statistically equivalent combined learning gains (p=.015). This establishes human-level performance at augmenting human capabilities through GRE tutoring for a broad mix of LLMs, including lightweight and open-weight models. Critically, the AI–human tutoring equivalence found here may represent a lower bound on LLM capabilities for augmenting humans because further gains in learning and cost-effectiveness are expected as models improve (Zhao et al., 2026; Gundlach

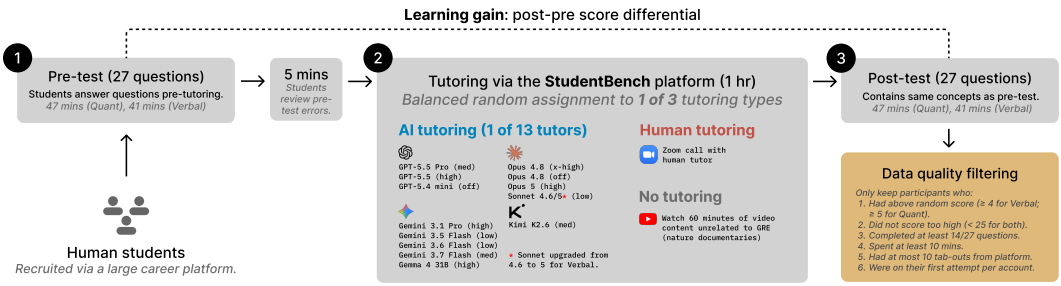

Figure 2: **Experimental design of the learning-gain study. 1,** Students completed either a 27-question Quantitative or Verbal **pre**-test, with eligibility and effort screening, and then spent five minutes reviewing incorrect answers. **2,** Students were assigned to one of three tutoring conditions: AI tutoring (with one of 13 LLM-based tutors; Table A.3), human tutoring, or no-tutoring control. **3,** After tutoring, students completed a **post**-test with the same length and concepts as the pre-test.

et al., 2025). For example, between May and September 2026, Gemini Flash's Terminal-Bench 2.1 score rose from 76% (Gemini 3.5 Flash) to 89% (Gemini 3.8 Flash) (Google DeepMind, 2026a;b), while Gemini 3.6 Flash's token prices fell 50% (Google, 2026). Among the AI tutors tested here, several have higher mean learning gains than human tutoring in Quantitative tests, whereas human tutoring retains the highest whole-section mean in Verbal tests (Figure D.2).

Our decision to study the conditions necessary for RHSI from a one-on-one learning perspective is motivated by the "two-sigma problem" of Bloom (1984), which asked how the benefits of one-on-one tutoring could be made broadly affordable. If LLMs can deliver effective tutoring at low cost, access to individualized instruction will be less limited by what a family can pay. StudentBench measures both learning and tutoring cost: we evaluate the cost to increase a student's GRE learning gain by one percentage point and discover markedly different cost-to-learning-gain ratios across AI tutors (Figures 4A and 5). Remarkably, one AI tutor achieved learning gains statistically equivalent to expert human tutoring (p=.044) at approximately $918\times$ lower cost per percentage point of learning gain. Its mean inference cost was \$0.067 for the entire tutoring session (lesson planning, practice-problem creation, and one hour of interactive tutoring).

To further understand LLM capabilities for augmenting human intelligence, we introduce three evaluations of AI teaching: (1) lesson planning, (2) practice-problem creation, and (3) conversational pedagogy. We conducted a second study, recruiting expert GRE tutors to compare LLM-generated lesson plans and practice problems through pairwise evaluations (Section 4.1). These evaluations produce AI tutor leaderboards from expert reviews of teaching materials and rule-based analysis of recorded tutoring conversations.

Finally, we examine how tutoring unfolds within student sessions. In the Quantitative section, faster AI replies were associated with more student engagement, more engagement with more correct practice, and more correct practice with larger learning gains (Appendix D.2).

Overall, StudentBench establishes that AI tutoring can produce GRE learning gains statistically equivalent to human tutoring, and introduces a new set of evaluation paradigms that measure real student learning gains on GRE-style assessments. To support future research, we have released the study data and reproduction code (links withheld for anonymous review).

2 Measuring Learning and Teaching

To evaluate how effectively AI tutors teach, we compared student learning gains across three conditions: AI tutoring, human tutoring, and no-tutoring control. In each session, students took a pre-test, spent one hour in their assigned condition, and then took a post-test with different questions (Figure 2). Each assessment matched the length and timing of the GRE exam: 27 questions in 47 minutes for Quantitative and 41 minutes for Verbal. After applying the data-quality filters shown in Figure 2, our analysis included 2,469 Quantitative and Verbal sessions from 2,383 students. Appendix A.3 shows the study interfaces in which students received AI or human tutoring and took assessments, and how experts compared lesson plans (Figures A.1–A.5).

2.1 STUDY SETUP

The pre-tests and post-tests for Quantitative and Verbal were crafted by former GRE exam creators from ETS and Kaplan, who designed new questions to match the difficulty, coverage, order, categories, and question types of prior GRE exams. Old GRE questions were never used unchanged because LLMs may have been trained on publicly available former GRE questions (Brown et al., 2020).

AI tutoring encompasses lesson planning, text-based conversational tutoring, and practice-problem generation. Students were randomly assigned a tutoring condition, and those assigned an AI model were not told which. Students assigned to the control condition watched educational videos unrelated to the GRE. Human tutors taught one-on-one in live video calls. Both human and AI tutors were given the student's pre-test mistakes to guide lesson planning and teaching, but were not given access to the questions on the post-test. Across all conditions, students spent about five minutes reviewing their mistakes on the graded pre-test.

Every AI tutor is built with two prompts for (1) lesson planning and problem creation, and (2) interactive tutoring live with the student. We use minimal software scaffolding and prompt adjustments (Appendix A.5, Figure G.1), so we can measure LLM teaching capabilities with limited influence from our design. The AI tutors determined what to teach, planned lessons, tutored students, and generated practice problems and answer keys.

Our results establish what AI tutors can achieve with low guidance; further prompt and harness improvements are likely to improve AI tutor performance. For example, on ARC-AGI-3, changes in harnesses and software scaffolding led to substantial performance gains (Karten et al., 2026).

The study took place over a two-month period in summer 2026. Through a large career platform, we invited approximately 70,000 randomly selected students across all majors, primarily aged 18 to 23. Within the AI condition, StudentBench randomly assigned students to different *AI tutors*, keeping counts balanced across available AI tutors. Each AI tutor is defined by its model and reasoning setting. New AI tutors entered the study as they became available. Our study was reviewed and approved by an independent IRB. Participants were paid $50 USD upon successfully completing a session.

2.2 MEASUREMENT AND METRICS

We define *learning gain* as post-test score minus pre-test score, in percentage points, where the score is the percentage of correctly answered questions. Unanswered questions count as incorrect, as in the GRE. To account for differences in average pre-test score across conditions when comparing learning gains, we use the analysis of covariance (ANCOVA) approach (Vickers & Altman, 2001). See Appendix B.1 for details. For AI $-$ human equivalence comparisons, we also adjust for assessment form and section, and account for shared tutors and repeated students using CR2 covariance with Satterthwaite degrees of freedom (Pustejovsky & Tipton, 2018). We test equivalence using standard two one-sided tests at α=.05 (Schuirmann, 1987; Lakens, 2017). We use equivalence bounds of ± 0.25 pooled standard deviations of learning gains, consistent with standardized bounds used in prior analyses of reading comprehension and treatment outcomes (Schwabe et al., 2022; Steinert et al., 2017). Equivalence requires the entire 90% confidence interval for the AI $-$ human difference to be within these bounds, corresponding to ± 4.09 percentage points for the combined GRE comparison. To evaluate human equivalence for individual (non-pooled) AI tutors, we use the same model and bounds in separate tests at α=.05, without correction for multiple comparisons across tutors.

2.3 EVALUATION WITH HUMAN EXPERTS

In a second study, 51 expert human tutors completed 2,028 pairwise reviews of AI-generated lesson plans and their embedded practice problems based on 381 student pre-tests. Each comparison paired one lesson plan with another generated by a different AI tutor from the same pre-test. For each pair, reviewers answered eight comparison questions: five about lesson planning, used for Figure 3A, and three about practice problems, used for Figure 3B. Appendix E describes the selected pre-tests, generation conditions, and individual criteria.

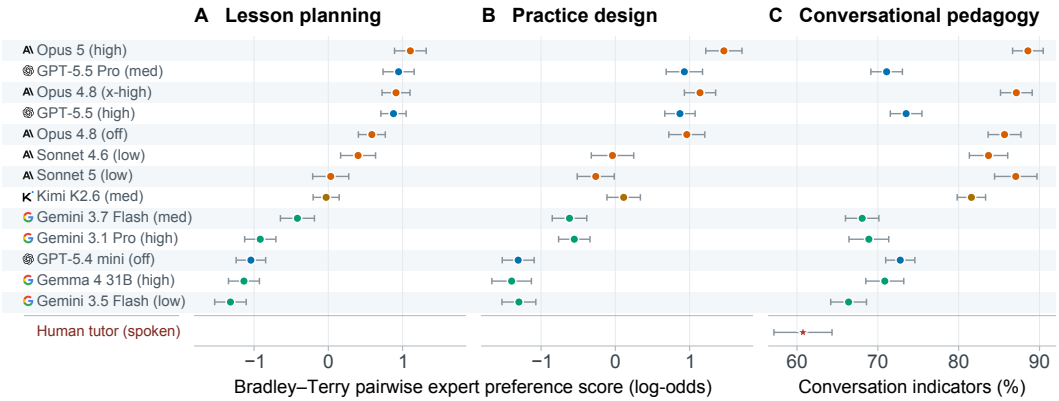

Figure 3: **Three evaluations of AI teaching.** Rows are ordered by lesson-planning score. **A, B,** Expert preferences for lesson planning and practice-problem creation, fitted with separate centered Bradley–Terry models; score differences are log-odds. Ratings marked as ties are excluded. Planning uses 9,577 ratings across 381 pre-tests; practice design uses 5,265 across 380. Whiskers are pointwise 95% intervals clustered by pre-test. **C,** Mean prevalence of six fixed conversation indicators across 1,971 AI and 135 human transcripts, with 95% Student-t intervals. The human reference is spoken tutoring; AI sessions use typed chat. Scores should be compared only within panels; blank cells mean not evaluated. Colors identify model families. See Appendices E and F for evaluation details.

2.4 Data quality and assessment integrity

Sessions were excluded for incompleteness, low effort, rapid responses, below-random performance, frequent tab switching, and pre-test scores too high for improvement ($> 24/27$). See Appendix A.2 for data filter details and Table A.2 for exclusion counts. To avoid reward hacking (Amodei et al., 2016), AI tutors were never given access to the post-test, and the prompts explicitly prohibited speculating about post-test content (Appendix A.5).

3 AI and Human Tutoring Yield Significant and Comparable Gains

Human tutoring and AI tutoring pooled across all AI tutors produced equivalent mean learning gains across the combined Quantitative and Verbal sessions (p=.015, $\pm$0.25-SD bounds; also equivalent under the stricter $\pm$0.20-SD bounds, p=.023; 1.59 d.f.). The combined AI $-$ human difference was -0.58 percentage points (90% CI $[-2.18, 1.03]$). At the $\pm$0.25-SD margin, equivalence was also established for Quantitative, but not Verbal.

Our pooled AI result spans 13 capability-varied AI tutors per section, including non-frontier and open-weight models. After adjustment for pre-test score, the AI $-$ control difference in learning gain was 6.86 percentage points in Quantitative (95% CI $[4.02, 9.69]$) and 5.47 in Verbal ($[2.46, 8.47]$). Gains with human tutoring were also higher than those in the control condition in both sections. Across both sections, the AI $-$ control difference was 6.15 percentage points ($[4.08, 8.21]$). That is about 1.5–2 more correct answers (out of 27) than in the control.

Figure 1A shows the frontier of AI tutoring across seven GRE domains. This visualization helps us see where the AI tutors stand relative to humans. In Quantitative, pooled AI and human mean learning gains are generally close, while the best AI tutors have higher mean gains than humans in three of the four domains. In Verbal, human mean gains are close to those of the best AI tutors, but remain above the pooled AI mean in all three domains. This contrast shows us where AI tutoring capabilities are strongest and where human tutors still have an advantage.

We were curious which AI tutors teach high-performing students best. To study this, we used item response theory to estimate proficiency on the Quantitative pre-test (Lord, 1980) and compared learning gains among students in the top 25%. These students have less room for improvement, yet four of the five highest mean gains came from Gemini tutors, led by Gemini 3.5 Flash. This finding may suggest a differentiated strength of the Gemini family of models in teaching high performers.

See Appendix A for tutor details; Appendix B for sensitivity checks; Appendix C for learning-gain estimates and starting-proficiency analyses; and Appendix D for learning gains by section.

4 THREE EVALUATIONS OF AI TEACHING

In this section, we present three evaluations of AI teaching. For each pair of AI lesson plans, expert reviewers answered eight comparison questions: five about lesson planning, used to build the leaderboard in Figure 3A, and three about practice problems, used to build the leaderboard in Figure 3B. We separate the two groups because they are distinct teaching capabilities. We also evaluate pedagogical characteristics of tutoring conversations to create the leaderboard in Figure 3C.

4.1 LESSON PLANNING AND PRACTICE-PROBLEM CREATION

The *lesson-planning leaderboard* evaluates concept relevance, grouping, prioritization, time allocation, and test-taking strategies (Anderson et al., 1995; Corbett & Anderson, 1995). The *practice-problem design leaderboard* evaluates alignment with the student's mistakes, appropriate difficulty, and accuracy of examples and answer keys. Appendix E gives the criteria and detailed rankings.

Across all three leaderboards, Anthropic models, particularly Opus, perform strongly. These models are preferred by expert tutors for lesson planning and practice-problem creation (Figure 3A, B) and more frequently exhibit the teaching behaviors we identified from prior research and evaluated using six transcript rules (Figure 3C). Satisfyingly, the leaderboards clearly distinguish models, with non-overlapping confidence intervals for most model pairs in Figure 3A, B.

Students could flag a practice answer with the button "I disagree with this answer — continue." GPT-5.4 mini had the highest observed flag rate: students disputed answers to approximately 11% of answered Quantitative practice problems. These flags record students' disagreements with the LLM-generated answers, not verified errors. Figure F.3 compares all AI tutor rates across the two sections and their combination. Interestingly, AI tutors scoring highly on practice-problem design (Figure 3B), i.e., AI tutors preferred by experts for practice-problem creation, also tended to receive fewer student answer disputes in Quantitative and combined analyses (Spearman ρ=0.71 and 0.77; p=.035 and p=.033; Appendix F.2).

4.2 CONVERSATIONAL PEDAGOGY

During a teaching session, human and AI tutors can ask students to explain (Aleven & Koedinger, 2002), adapt their help to a concrete mistake, or leave room for an attempt before demonstrating a solution (Wood et al., 1976; Koedinger & Aleven, 2007). Each action elicits different levels of cognitive engagement from the student (Chi, 2009). We created six fixed text and turn-order rules to measure the prevalence of these elements in the tutoring sessions. For each transcript, we score each of the six rules as 1 if the element shows up at least once and 0 otherwise, then average these six values. A tutor's score is the mean of these transcript averages, multiplied by 100. Note this gives a bias towards sessions with more turns or more verbose AI tutors. Table F.1 overviews the six rules, drawing from established literature in the learning and cognitive sciences (VanLehn, 2011; Chi et al., 1994; Sweller & Cooper, 1985; Dunlosky et al., 2013; Renkl, 2014).

Strikingly, in Figure 3C, AI tutors cluster by model family, with no overlap between family ranges (similarly colored points grouped themselves into distinct vertical bands). This suggests that pedagogical characteristics of AI tutors may reflect company-wide training practices shared across a provider's models. Individual indicators and section-specific results are in Figures F.1 and F.2.

5 AI TUTORING COST, LATENCY, AND ENGAGEMENT

In this section, we explore the Pareto frontier of learning gain, cost, and latency, and how latency relates to student engagement. For each session, analysis is computed directly from recorded session data. We combine Quantitative and Verbal results in the main text and separate results in Appendix D.

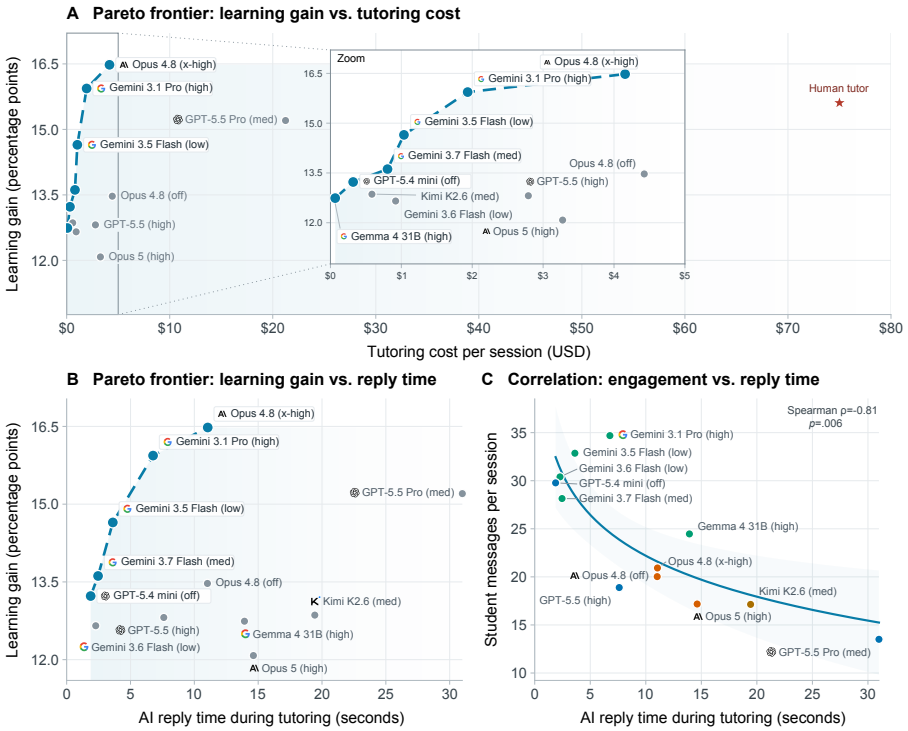

Figure 4: **Tutoring cost, AI reply time, and student engagement. A, B,** Mean learning gain versus (A) total AI cost per session and (B) median AI reply time. Blue marks the Pareto frontier. The red star uses a $75-per-hour human reference. **C,** Mean student-authored chat messages per session versus the median reply time. Shading shows the pointwise 95% confidence band.

5.1 Cost of Learning

Figure 4A shows which AI models provide the highest learning gain at the lowest cost. The AI tutoring cost depicted is the total of recorded or reconstructed AI inference costs (97% coverage of API calls) in the session across lesson planning, interactive chat tutoring, and practice-problem generation. We use $75 per hour as the market reference rate for human tutoring, based on a published rate (Jantzi Test Prep, 2018). Appendix D.1 gives technical details.

In Figure 4A, we see that Gemini 3.5 Flash had a similar learning gain to GPT-5.5 Pro while costing 20× less. All AI tutors on the Pareto frontier had a mean cost per session of less than $5, suggesting that *AI tutors can democratize access to high-quality learning*. Across the 12 AI tutors, mean inference cost per session ranged from $0.067 for Gemma 4 31B to $21.24 for GPT-5.5 Pro. We can use Figure 4 to measure where AI tutors strictly Pareto-dominate others: for example, Gemini 3.1 Pro had higher mean learning gain, lower inference cost, and faster replies than GPT-5.5 Pro.

5.2 Faster Replies are Associated with More Student Dialogue

Figure 4C depicts how AI reply time[1] correlates with student engagement.[2] Across the 12 AI tutors, lower AI latency was strongly associated with higher student engagement (Spearman ρ=−0.81, p=.006). Across AI tutors, the association holds for Quantitative and Verbal separately (Figure D.3). In an exploratory analysis of 1,137 Quantitative AI-tutoring sessions, faster replies were associated with more student engagement, more engagement with more correct practice,[3] and more correct practice with greater learning gains (all $p \leq .002$), linking AI latency to learning gain.

[1] *Reply time (latency)* measures how long the AI model takes to return a complete response.

[2] *Engagement* counts non-blank student chat messages per session, excluding practice-answer submissions.

[3] *Correct practice* counts distinct closed-response problems marked correct on the first attempt using the AI tutor's answer key.

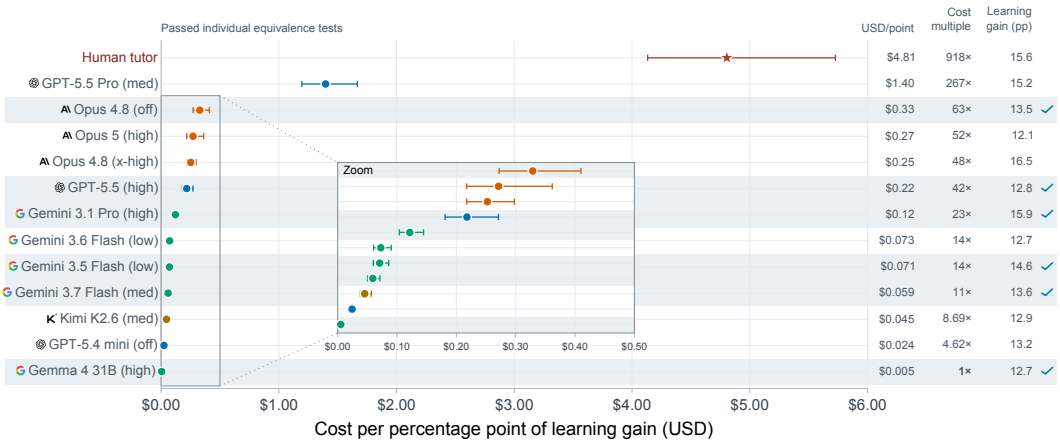

Figure 5: **Cost per percentage point of learning gain.** Cost per percentage point (pp) is the mean session cost divided by the mean learning gain. Whiskers show 95% confidence intervals. AI costs reflect inference; human tutoring uses a $75-per-hour reference. Cost multiples use Gemma 4 31B as the $1\times$ reference and are calculated from unrounded values. Blue shading and checkmarks identify all six AI tutors that passed individual equivalence tests against human tutoring ($p<.05$). See Figure D.1 for section-specific rankings.

6 The Cost per Percentage Point of Learning Gain

One direction we found particularly satisfying with StudentBench is measuring learning gains relative to cost. Section 5.1 compares session costs. Here, we propose a standardized reporting unit: *USD per percentage point of learning gain*, calculated as mean session cost divided by mean learning gain. This unit gives us a common basis to compare AI and human tutoring costs within this study. We measure only immediate GRE learning gains, but this approach may help quantify the cost of augmenting human intelligence more broadly.

We test whether each AI tutor's learning gains are equivalent to human tutoring, and find that Gemma 4 31B achieved GRE learning gains equivalent to expert human tutoring ($p=.044$) at $918\times$ lower cost per percentage point of learning gain. This has major implications for the democratization of opportunities for students. We set the lowest-cost tutor, Gemma 4 31B, as the $1\times$ cost reference and express each tutor's cost per percentage point relative to it. Figure 5 shows the resulting comparison and highlights the six AI tutors that passed individual equivalence tests against human tutoring. See Appendix B.7 for the tests and calculations.

7 Background and Related Work

Teaching technologies and human capabilities. The pursuit of teaching and human-augmenting technology spans millennia, from Mesopotamian lexical tablets and hornbooks (Lecompte & Naccaro, 2025; Plimpton, 1916) to Pressey's teaching machine (Pressey, 1928), and it shaped computing, including early visions of human–computer symbiosis (Licklider, 1960), the co-design of tools and training to expand human problem-solving capacity (Engelbart, 1962), and distinctions between *effects with* and *effects of* technology (Salomon et al., 1991).[4]

Bloom's two-sigma problem, how to scale the benefits of individual tutoring (Bloom, 1984), motivated intelligent tutoring systems that model student knowledge and personalize instruction, including cognitive tutors (Anderson et al., 1995), ACT-R (Anderson et al., 2004), knowledge tracing (Corbett & Anderson, 1995), and AutoTutor (Graesser et al., 2005). Subsequent reviews found that some systems achieved learning outcomes comparable to human tutoring (VanLehn, 2011; Ma et al., 2014). StudentBench studies whether general-purpose LLMs with minimal scaffolding can also produce substantial learning gains.

[4]Teaching technology should develop capabilities that are retained when its assistance is removed, as depicted in *The Matrix* (1999): Neo permanently acquires martial-arts skills after a brief computer interaction.

Students' performance with and without guidance can reveal their zone of proximal development (Vygotsky, 1978), and cognitive tutors that elicit explanations can improve learning (Aleven & Koedinger, 2002). StudentBench examines related aspects of tutoring dialogue: scaffolding cues (Wood et al., 1976), requests for explanations and reasoning checks (Aleven & Koedinger, 2002), interactive questioning (Graesser et al., 2005), and the balance between providing help and leaving room for students to reason (Koedinger & Aleven, 2007). Effective tutors must also keep students motivated and adapt to learning difficulties. Research on computer tutoring shows that confusion, engagement, frustration, and boredom can change within a single learning session (D'Mello & Graesser, 2012). *Self-determination theory* emphasizes autonomy, competence, and relatedness as drivers of motivation (Ryan & Deci, 2000). Student activity alone does not imply learning, as students can progress through hints and feedback without engaging in the intended reasoning (Baker et al., 2004), and completing more practice problems does not necessarily imply greater retention (Rohrer & Taylor, 2006). StudentBench therefore relates conversational pedagogy, student engagement, and practice activity during tutoring to learning gains measured afterward.

AI tutoring and benchmarks. Pardos & Bhandari (2024) compared AI-generated and human-authored mathematics hints; StudentBench extends the comparison to live tutoring conversations. Kestin et al. (2025) compared a structured AI physics tutor with in-class active learning, whereas we study individual human tutoring. While Wang et al. (2025) study AI systems to assist human tutors and LearnLM Team, Google and Eedi (2025) study AI supervised by them, StudentBench evaluates autonomous AI tutors. StudentBench measures learning with unaided post-tests, as Bastani et al. (2025) found that unguarded AI help improved practice but reduced later unaided performance in mathematics. Contractor & Reyes (2026) found gains after AI-assisted study and writing.

Few LLM benchmarks evaluate human augmentation: TutorBench evaluates model responses to expert-authored student-persona conversations (Srinivasa et al., 2025), while INSIDE simulates students' code submissions and reasoning (Niousha et al., 2026). In contrast, StudentBench measures real students' learning gains after live tutoring. Its expert ratings of lesson plans and practice design complement psychometric evaluations of AI-generated exam questions (Isley et al., 2026).

8 Limitations

Longitudinal retention studies take longer to release, and rapid increases in AI capabilities (Zhao et al., 2026) can outdate the AI tutors faster than we can publish meaningful results. One limitation of this study is that we measured immediate learning gains without evaluating whether they persist over months. A strength of this work is the large sample size; maintaining a large longitudinal sample would increase recruitment and follow-up costs. Other limitations exist due to the study population: participants were all adults able to read and write in English, who had access to the recruiting platform. The study did not test deployment across languages, devices, or educational settings. The leaderboard evaluations in Figure 3 characterize teaching quality based on expert tutor evaluations and observed conversational behaviors, not based on how much learning occurs. We did not include a control condition in which students worked through GRE practice problems on their own. Our comparison therefore does not separately measure the added benefit of interacting with an AI tutor over doing practice problems alone.

9 Conclusion

Pooled AI and expert human tutoring produced statistically equivalent combined GRE learning gains. Among the six AI tutors that passed individual equivalence tests, Gemma 4 31B achieved these gains at $918\times$ lower cost per percentage point than human tutoring. These findings support the teaching condition we proposed for RHSI. Together with evidence that tutoring systems can improve learning across subjects (Ma et al., 2014; Kestin et al., 2025), our findings suggest that AI tutoring may also help students prepare for other standardized exams, such as the SAT and MCAT. StudentBench also introduces five benchmark leaderboards that separate AI tutors across lesson planning, practice-problem creation, conversational pedagogy, cost, and engagement.

Lastly, StudentBench contributes infrastructure and data collection methods demonstrating a new evaluation paradigm that measures how AI can *uplift, not replace*, human intelligence.

ETHICS STATEMENT

The StudentBench study protocol was approved by an independent Institutional Review Board (protocol number withheld for anonymous review). All participants consented to research use. The public release uses anonymous session identifiers and excludes identity mappings, individual locations, and audio/video recordings. The study uses research assessments and does not alter participants' official examination records or admissions decisions.

REPRODUCIBILITY STATEMENT

Data and code have been publicly released with the preprint. The release identifies the authors, so links are omitted for double-blind review. Further analysis details are available in Appendices B and D.1.

LLM USAGE STATEMENT

We used large language models to improve the writing of this paper, including grammar, wording, and clarity. All text was reviewed and edited by the authors, who take full responsibility for the content of the paper.

REFERENCES

Vincent A.W.M.M. Aleven and Kenneth R. Koedinger. An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2):147–179, 2002. doi: 10.1207/s15516709cog2602_1. URL https://doi.org/10.1207/s15516709cog2602_1.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, 2016. URL https://arxiv.org/abs/1606.06565. arXiv preprint arXiv:1606.06565.

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. URL https://doi.org/10.1207/s15327809jls0402_2.

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An Integrated Theory of the Mind. *Psychological Review*, 111(4):1036–1060, 2004. doi: 10.1037/0033-295x.111.4.1036. URL https://doi.org/10.1037/0033-295x.111.4.1036.

Anthropic. System card: Claude Fable 5.1 & Claude Mythos 5.1. Technical report, Anthropic, September 2026. URL https://www.anthropic.com/claude-fable-5-1-mythos-5-1-system-card. Published 1 September 2026.

Artificial Analysis. Announcing the Artificial Analysis Intelligence Index v4.3, September 2026. URL https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-3. Published 7 September 2026; accessed 12 September 2026.

Ryan Shaun Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 383–390, 2004. doi: 10.1145/985692.985741. URL https://doi.org/10.1145/985692.985741.

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. URL https://doi.org/10.1073/pnas.2422633122.

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. doi: 10.3102/0013189x013006004. URL https://doi.org/10.3102/0013189x013006004.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249, 2011. doi: 10.1198/jbes.2010.07136.

Michelene T. H. Chi. Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. *Topics in Cognitive Science*, 1(1):73–105, 2009. doi: 10.1111/j.1756-8765.2008.01005.x. URL https://doi.org/10.1111/j.1756-8765.2008.01005.x.

Michelene T. H. Chi and Ruth Wylie. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243, 2014. doi: 10.1080/00461520.2014.965823. URL https://doi.org/10.1080/00461520.2014.965823.

Michelene T. H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477, 1994. doi: 10.1207/s15516709cog1803_3. URL https://doi.org/10.1207/s15516709cog1803_3.

College Board. *Assessment Framework for the Digital SAT Suite*. College Board, New York, August 2024. URL https://satsuite.collegeboard.org/media/pdf/assessment-framework-for-digital-sat-suite.pdf. Version 3.01.

Zara Contractor and Germán Reyes. Experimental evidence on the learning impact of generative AI, 2026. URL https://arxiv.org/abs/2607.08849. arXiv preprint arXiv:2607.08849.

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, 1995. doi: 10.1007/bf01099821. URL https://doi.org/10.1007/bf01099821.

Sidney D'Mello and Art Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012. doi: 10.1016/j.learninstruc.2011.10.001. URL https://doi.org/10.1016/j.learninstruc.2011.10.001.

John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1):4–58, 2013. doi: 10.1177/1529100612453266. URL https://doi.org/10.1177/1529100612453266.

Educational Testing Service. *GRE Guide to the Use of Scores*. Educational Testing Service, 2026. URL https://www.ets.org/content/dam/ets-org/pdfs/gre/gre-guide-to-the-use-of-scores.pdf. 2026–27 edition.

Educational Testing Service. Overview of the quantitative reasoning measure. GRE General Test official content specification, n.d. URL https://www.ets.org/gre/test-takers/general-test/prepare/content/quantitative-reasoning.html. Accessed 12 September 2026.

Douglas C. Engelbart. Augmenting human intellect: A conceptual framework. Technical Report AFOSR-3223, Stanford Research Institute, Menlo Park, CA, October 1962. URL https://dougengelbart.org/content/view/138/.

Irving John Good. Speculations concerning the first ultraintelligent machine. In Franz L. Alt and Morris Rubinoff (eds.), *Advances in Computers*, volume 6, pp. 31–88. Academic Press, 1965. doi: 10.1016/S0065-2458(08)60418-0. URL https://www.sciencedirect.com/science/article/pii/S0065245808604180.

Google. Gemini Developer API pricing. https://ai.google.dev/gemini-api/docs/pricing, 2026. Accessed September 22, 2026; Gemini 3.6 Flash standard paid-tier promotional rates through December 31, 2026.

Google DeepMind. Gemini 3.5 Flash model card. Technical report, Google DeepMind, May 2026a. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf.

Google DeepMind. Gemini 3.8 Flash model card. Technical report, Google DeepMind, September 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. September 2026.

Arthur C. Graesser, Patrick Chipman, Brian C. Haynes, and Andrew Olney. AutoTutor: An Intelligent Tutoring System With Mixed-Initiative Dialogue. *IEEE Transactions on Education*, 48(4):612–618, 2005. doi: 10.1109/te.2005.856149. URL https://doi.org/10.1109/te.2005.856149.

Hans Gundlach, Jayson Lynch, Matthias Mertens, and Neil Thompson. The price of progress: Price performance and the future of AI, 2025. URL https://arxiv.org/abs/2511.23455v2. Version 2, revised 23 March 2026.

Calvin Isley, Joshua Gilbert, Evangelos Kassos, Michaela Kocher, Allen Nie, Emma Brunskill, Ben Domingue, Jake Hofman, Joscha Legewie, Teddy Svoronos, Charlotte Tuminelli, and Sharad Goel. Assessing the quality of AI-generated exams: A large-scale field study. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45):38626–38634, 2026. doi: 10.1609/aaai.v40i45.41205. URL https://ojs.aaai.org/index.php/AAAI/article/view/41205.

Jantzi Test Prep. GRE prep tutoring, 2018. URL https://4satprep.com/private-tutoring/gre-prep-tutoring/. Expert tier: $75 per hour; tutors have at least 100 hours of tutoring experience. Accessed 21 September 2026.

Manu Kapur. Productive Failure. *Cognition and Instruction*, 26(3):379–424, 2008. doi: 10.1080/07370000802212669. URL https://doi.org/10.1080/07370000802212669.

Seth Karten, Alex L. Zhang, Kevin Thomas, Sebastian Müller, Elie Bakouch, Daniel Auras, Mika Senghaas, Fares Obeid, Konstantin Dunas, Johannes Hagemann, and Sami Jaghouar. Prime Agent: A self-improving RLM harness. *arXiv preprint arXiv:2608.23552*, 2026. doi: 10.48550/arXiv.2608.23552. URL https://arxiv.org/abs/2608.23552.

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. URL https://www.nature.com/articles/s41598-025-97652-6.

Kenneth R. Koedinger and Vincent Aleven. Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review*, 19(3):239–264, 2007. doi: 10.1007/s10648-007-9049-0. URL https://doi.org/10.1007/s10648-007-9049-0.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177. URL https://journals.sagepub.com/doi/10.1177/1948550617697177.

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

Camille Lecompte and Hugo Naccaro. The colophons of the archaic lists and their social setting. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 115(1):1–35, 2025. doi: 10.1515/za-2025-0001. URL https://doi.org/10.1515/za-2025-0001.

J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1):4–11, March 1960. doi: 10.1109/THFE2.1960.4503259. URL https://www.cl.cam.ac.uk/~pr10/iui/licklider60.pdf.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

Wenting Ma, Olusola O. Adesope, John C. Nesbit, and Qing Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918, 2014. doi: 10.1037/a0037123. URL https://doi.org/10.1037/a0037123.

James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985. doi: 10.1016/0304-4076(85)90158-7. URL https://www.sciencedirect.com/science/article/pii/0304407685901587.

Casey W. Miller, Benjamin M. Zwickl, Julie R. Posselt, Rachel T. Silvestrini, and Theodore Hodapp. Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion. *Science Advances*, 5(1):eaat7550, 2019. doi: 10.1126/sciadv.aat7550. URL https://doi.org/10.1126/sciadv.aat7550.

Rose Niousha, Minwoo Kang, and Narges Norouzi. INSIDE the student's mind: Jointly modeling latent reasoning and action in LLM student simulators, 2026. URL https://arxiv.org/abs/2608.10492. Accepted at the Conference on Language Modeling (COLM), 2026; arXiv:2608.10492.

OpenAI. GPT-6 Astra system card. Technical report, OpenAI, September 2026. URL https://deploymentsafety.openai.com/gpt-6-astra. Published 3 September 2026; revised 9 September 2026.

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. doi: 10.1371/journal.pone.0304013. URL https://doi.org/10.1371/journal.pone.0304013.

George Arthur Plimpton. The hornbook and its use in America. *Proceedings of the American Antiquarian Society*, 26(2):264–272, October 1916. URL https://www.americanantiquarian.org/node/10682.

Sidney L. Pressey. Machine for intelligence tests. U.S. Patent 1,670,480, May 1928. URL https://patents.google.com/patent/US1670480A/en. Filed 30 January 1926; granted 22 May 1928.

James E. Pustejovsky and Elizabeth Tipton. Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, 2018. doi: 10.1080/07350015.2016.1247004.

Alexander Renkl. Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, 38(1):1–37, 2014. doi: 10.1111/cogs.12086. URL https://doi.org/10.1111/cogs.12086.

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. URL https://doi.org/10.1002/acp.1266.

Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000. doi: 10.1037/0003-066x.55.1.68. URL https://doi.org/10.1037/0003-066x.55.1.68.

Gavriel Salomon, David N. Perkins, and Tamar Globerson. Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher*, 20(3):2–9, April 1991. doi: 10.3102/0013189X020003002. URL https://doi.org/10.3102/0013189X020003002.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419. URL https://link.springer.com/article/10.1007/BF01068419.

Annika Schwabe, Fabienne Lind, Lukas Kosch, and Hajo G. Boomgaarden. No negative effects of reading on screen on comprehension of narrative texts compared to print: A meta-analysis. *Media Psychology*, 25(6):779–796, 2022. doi: 10.1080/15213269.2022.2070216. URL https://doi.org/10.1080/15213269.2022.2070216.

SpaceXAI. Model card: Grok 4.6. Technical report, xAI LLC, August 2026. URL https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf. Published 12 August 2026; revised 17 August 2026.

Rakshith S Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, Ed-Yeremai Hernandez Cardona, Anisha Gunjal, Yunzhong He, Bing Liu, and Chen Xing. TutorBench: A benchmark to assess tutoring capabilities of large language models, 2025. URL https://arxiv.org/abs/2510.02663. arXiv preprint arXiv:2510.02663.

Christiane Steinert, Thomas Munder, Sven Rabung, Jürgen Hoyer, and Falk Leichsenring. Psychodynamic therapy: As efficacious as other empirically supported treatments? A meta-analysis testing equivalence of outcomes. *American Journal of Psychiatry*, 174(10):943–953, 2017. doi: 10.1176/appi.ajp.2017.17010057. URL https://doi.org/10.1176/appi.ajp.2017.17010057.

John Sweller and Graham A Cooper. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1):59–89, 1985. doi: 10.1207/s1532690xci0201_3. URL https://doi.org/10.1207/s1532690xci0201_3.

Heather Turner and David Firth. Bradley–Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, 48(9):1–21, 2012. doi: 10.18637/jss.v048.i09. URL https://www.jstatsoft.org/article/view/v048i09.

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. URL https://doi.org/10.1080/00461520.2011.611369.

Andrew J. Vickers and Douglas G. Altman. Analysing controlled trials with baseline and follow up measurements. *BMJ*, 323(7321):1123–1124, 2001. doi: 10.1136/bmj.323.7321.1123. URL https://www.bmj.com/content/323/7321/1123.

Vernor Vinge. The coming technological singularity: How to survive in the post-human era. In *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, pp. 11–22. National Aeronautics and Space Administration, 1993. URL https://ntrs.nasa.gov/citations/19940022856.

L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978. doi: 10.2307/j.ctvjf9vz4. URL https://www.jstor.org/stable/j.ctvjf9vz4. Edited by Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman.

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, 2025. URL https://arxiv.org/abs/2410.03017. arXiv:2410.03017, version 2; first posted 2024.

David Wood, Jerome S. Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976. doi: 10.1111/j.1469-7610.1976.tb00381.x. URL https://doi.org/10.1111/j.1469-7610.1976.tb00381.x.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel Machine: Open-ended evolution of self-improving agents. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=pUpzQZTvGY.

Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents, 2026b. URL https://arxiv.org/abs/2603.19461. arXiv preprint arXiv:2603.19461.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Yiwen Hu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *Frontiers of Computer Science*, 20: 2012627, 2026. doi: 10.1007/s11704-026-60308-3. URL https://doi.org/10.1007/s11704-026-60308-3.

APPENDIX

A STUDY PROTOCOL AND POPULATION

A.1 POPULATION AND STUDY CONDITIONS

In Section 2 we mentioned that students were recruited through a large career platform. Invitations went to over 70,000 randomly selected students across different majors. Participants were offered $50 for completing a session, independent of test performance. In total, the analysis includes 2,383 unique adults aged 18–67 (median 21); 83.3% were ages 18–24. Of these, 1,205 completed Quantitative only, 1,092 Verbal only, and 86 completed both. The Quantitative and Verbal analyses include 1,291 and 1,178 sessions, respectively. The human comparison included nine tutors per section, with 1–17 students per tutor in Quantitative and 1–21 in Verbal. Human tutors were former ETS or Kaplan employees responsible for writing GRE questions, or had at least five years of GRE tutoring experience. No live human tutor was allowed to see the post-test questions for any student they tutored. Human tutors were asked to focus on pre-test mistakes, order concepts by importance, and allocate the teaching time. Tutors could add notes about misconceptions and examples in the planning interface.

For Quantitative and Verbal, we created two GRE tests denoted P and Q. As stated in Section 2, the assessments were crafted by former GRE exam creators from ETS and Kaplan, who designed new questions to match the difficulty, coverage, order, categories, and question types of prior GRE exams whilst avoiding old exam questions to mitigate bias from training-data contamination. About half of the students are given P as their pre-test and Q as their post-test, while the others are given Q as their pre-test and P as their post-test to ensure the assessment-form order does not impact the results; thus giving an unbiased estimate of the learning gains. P→Q and Q→P denote the pre-test and post-test form order.

Each Quantitative test allows 47 minutes, and each Verbal test allows 41 minutes. Together, the two assessments and one-hour assigned condition allow up to 154 minutes for Quantitative and 142 minutes for Verbal. Students may finish the pre- or post-test early, but must complete the one-hour assigned condition.

Table A.1 reports starting scores and assessment-form order for the comparisons in Figure 1 and the study design in Section 2.

StudentBench used *least-fill assignment* for AI tutoring: each student was randomly assigned among the available AI tutors with the fewest assigned students. Initially, it counted all enrolled students; later, it counted only students who had completed or were still completing a session.

Every AI session in the main study used the same lesson-planning and tutoring prompts for its GRE section. Separate prompt comparisons are reported in Appendix G; data points from earlier control conditions are excluded from the main study.

We used the same process to develop the Quantitative and Verbal assessments. Former ETS GRE exam creators and GRE tutors with at least five years of experience used original GRE exams as a starting point, edited and created new questions, and adjusted their difficulty to match the actual GRE. No original GRE question was reused unchanged. An independent group of GRE test question creators reviewed the questions for difficulty. We replaced seven Quantitative question pairs with harder questions written and reviewed by GRE tutors.

Before conducting the study, we worked with expert GRE tutors to match every pre-test question to a corresponding post-test question with the same difficulty label, format, and subtopic.

A.2 DATA QUALITY

The learning comparisons in Figure 1 include only students who completed the study and met eligibility and participation requirements. The analysis excludes students with insufficient assessment effort, incomplete sessions, repeated attempts or documented deviations from the assigned condition. Pre-test score limits leave room for improvement while excluding scores below the study's minimum eligibility threshold. No filter uses learning gain, post-test score, or post-test correctness.

Table A.1: Starting scores and assessment-form order for the learning comparisons in Figure 1 and the study design in Section 2. Both forms were used as pre-tests and post-tests; the counts show how often each order was used within each condition. Pre-test scores are percentages correct; parentheses give the sample standard deviation. P→Q and Q→P denote the pre-test and post-test form order. The human tutoring condition had lower mean pre-test scores in both sections by chance; assignment was random and independent of pre-test score.

Section	Condition	Sessions	Pre-test, mean (SD)	P→Q	Q→P
Quantitative	AI	1,139	50.21 (21.16)	580	559
Quantitative	Human	61	45.17 (18.23)	27	34
Quantitative	Control	91	52.99 (21.22)	45	46
Verbal	AI	1,000	46.35 (19.47)	499	501
Verbal	Human	79	41.44 (15.43)	57	22
Verbal	Control	99	47.29 (19.08)	50	49

Eligibility and assessment effort. Students must consent to research use, be adults, be able to read and write in English, and be taking the assigned assessment for the first time. Records created for administration, testing, demonstrations, or staff quality checks are ineligible. Eligible pre-test scores are 5–24 correct answers in Quantitative and 4–24 in Verbal, out of 27. The lower limits reflect section-specific random-guessing baselines, which depend on the mix of question types and whether answer choices are provided. Both assessments must be completed, with at least 14 non-blank answers, at least ten minutes of recorded time on the questions, and no more than ten recorded departures from the assessment page. A page departure that triggers both a focus event and a visibility event counts once. The assessment interface prohibits outside assistance and records tab switching (Figure A.2); these records help identify departures from the protocol but cannot rule out outside assistance.

Participation in the assigned condition. AI sessions require at least three student messages, five practice-answer attempts and three distinct attempted practice problems. These thresholds require students to participate in the tutoring shown in Figure A.1; they do not depend on learning gain. Sessions are excluded for documented use of a lesson plan from another AI tutor, substantial additional treatment, or an incorrect study condition. The control condition requires students to spend at least five minutes reviewing the graded pre-test before watching the assigned videos. We report alternative prompt, control and pilot conditions separately from sessions that failed to follow the final protocol.

Rapid submissions. For each session, StudentBench monitors how quickly students submit answers on each assessment. The effort screen excludes sessions when, on either assessment, about 90% or more of consecutive answers are submitted less than ten seconds apart (90th percentile below ten seconds). It uses each question's first server-recorded submission and requires at least 13 intervals. The final five minutes are omitted so that rapid answers near the deadline do not alone trigger the screen. Insufficient timing data alone does not exclude a session.

Excluded records. Table A.2 combines the eligibility and protocol exclusions with the final adherence screen.

Table A.2: Data-quality exclusions for the study in Section 2 and the learning results in Figure 1. Each record has one primary reason within its screening stage. No record appears in both stages.

Primary exclusion reason	Quant	Verbal	Total
Eligibility and study-protocol checks			
Pre-test score below eligibility range	225	83	308
Pre-test score above eligibility range	266	38	304
Insufficient assessment effort	143	124	267
Incomplete study	139	34	173
Assessment score unavailable	87	51	138
Repeated attempt	11	15	26
Substantive additional treatment	2	1	3
Unexpected AI tutoring	1	1	2
Lesson plan from another AI tutor	0	1	1
Final adherence checks			
Insufficient AI tutoring participation	82	104	186
Incomplete assessment	23	6	29
Rapid assessment submissions	2	2	4

A.3 Study interfaces

The interfaces implement the study design in Section 2 and the expert evaluations in Section 4.1. Figures A.1–A.5 show how students took assessments and received AI or human tutoring, and how experts compared lesson plans.

A.4 Deployed AI tutors

Each GRE section tested 13 AI tutors[5] based on 12 models (Table A.3), with Opus 4.8 tested at two reasoning settings: off and extra-high. In figures and tables, parentheses in AI tutor names indicate the configured reasoning level or mode.

Kimi K2.6 requested `reasoning.effort=medium` through OpenRouter. Figures and tables use "off" for disabled reasoning: GPT-5.4 mini sent `reasoning_effort=none`, while Opus 4.8 omitted `thinking` and sent `output_config.effort=low`.

Table A.3: AI tutors with reasoning effort in parentheses, API endpoints, and Quantitative and Verbal sample sizes for the learning study. See Figure 1 for learning results.

AI tutor	Endpoint	Quant	Verbal
Gemini 3.1 Pro (high)	`gemini-3.1-pro-preview`	91	74
Gemini 3.5 Flash (low)	`gemini-3.5-flash`	97	80
Gemini 3.6 Flash (low)	`gemini-3.6-flash`	89	79
Gemini 3.7 Flash (med)	`gemini-3.7-flash`	87	92
Gemma 4 31B (high)	`gemma-4-31b-it`	87	72
GPT-5.4 mini (off)	`gpt-5.4-mini`	90	71
GPT-5.5 (high)	`gpt-5.5`	89	72
GPT-5.5 Pro (med)	`gpt-5.5-pro`	86	76
Kimi K2.6 (med)	`moonshotai/kimi-k2.6`	85	76
Opus 4.8 (off)	`claude-opus-4-8`	84	73
Opus 4.8 (x-high)	`claude-opus-4-8`	86	72
Opus 5 (high)	`claude-opus-5`	75	78
Sonnet (low)	`claude-sonnet-4-6`	93	0
	`claude-sonnet-5`	0	85

[5] Sonnet used version 4.6 in Quantitative and version 5 in Verbal.

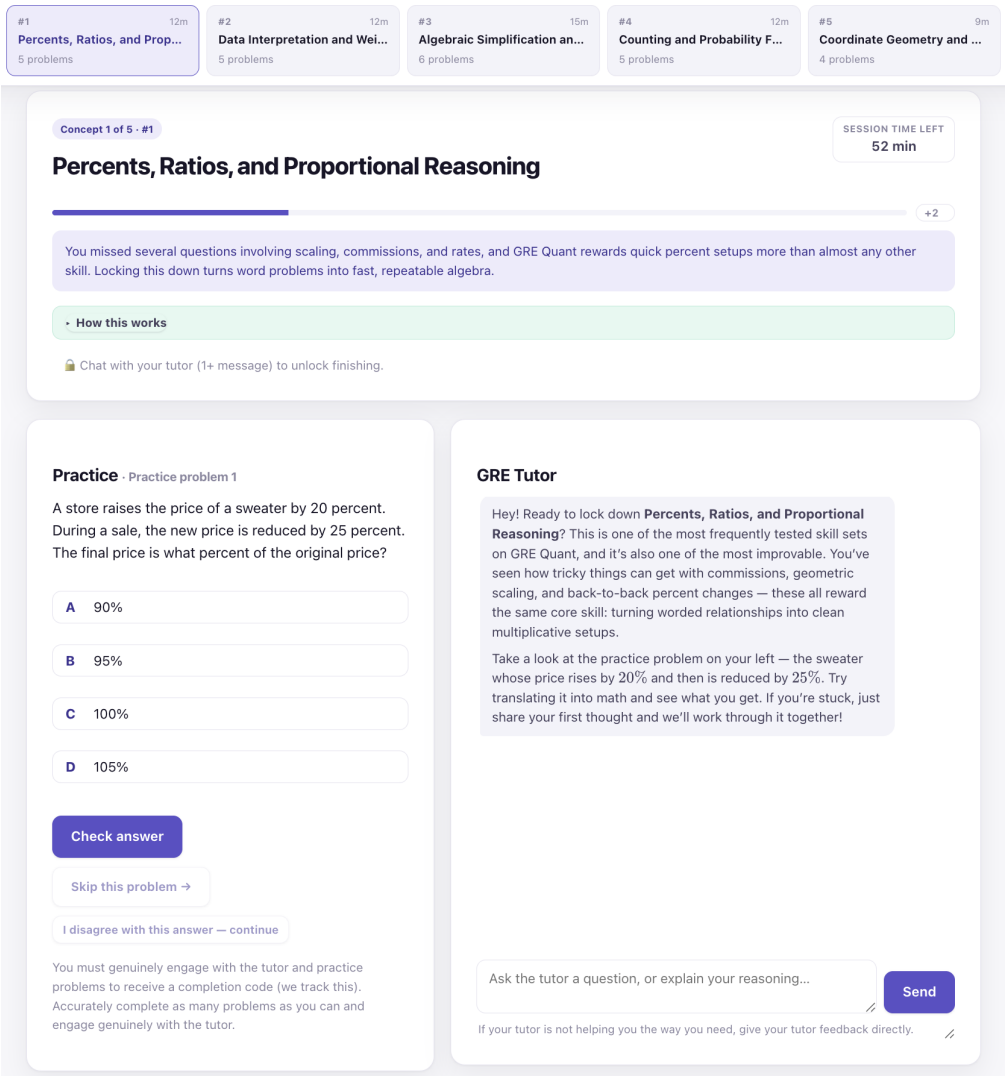

Figure A.1: **Student tutoring interface.** Lesson planning generates the concepts and time allocations across the top and the initial practice problems on the left. The live tutoring conversation appears on the right, with access to the student's practice answers. See Figure 2 for the study design.

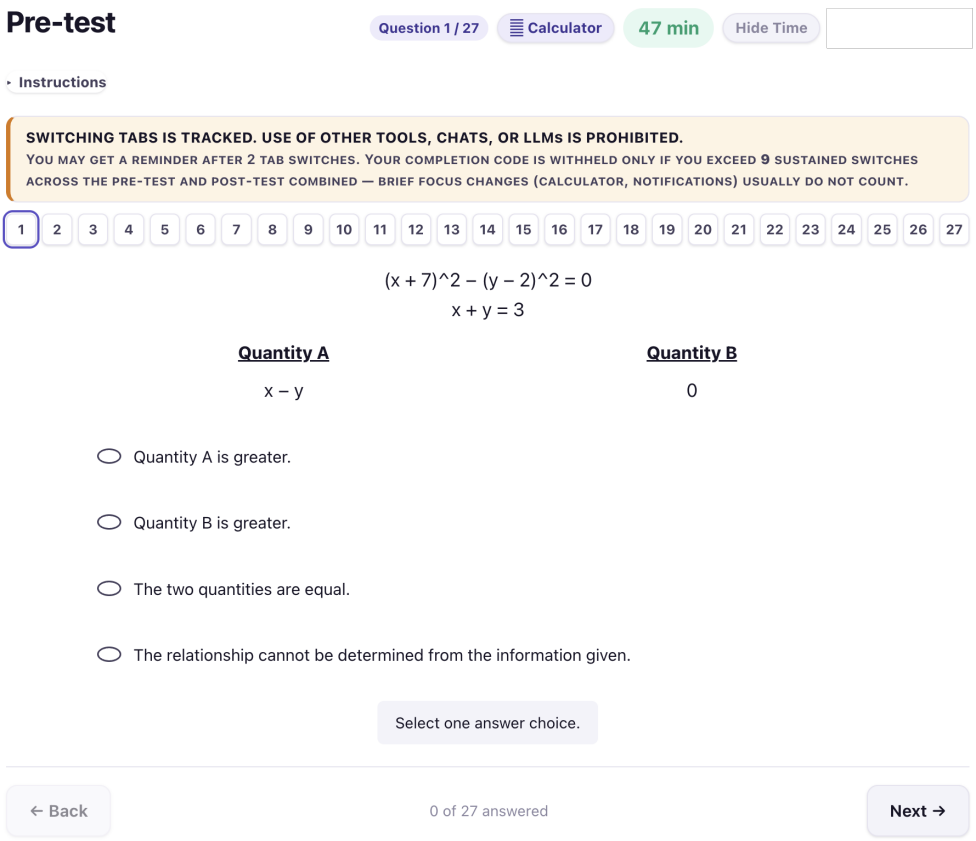

Figure A.2: **Quantitative assessment interface.** Pre-tests and post-tests use the same interface to measure unaided performance. The notice states that outside assistance is prohibited and tab switching is recorded. See Figure 1 for learning results.

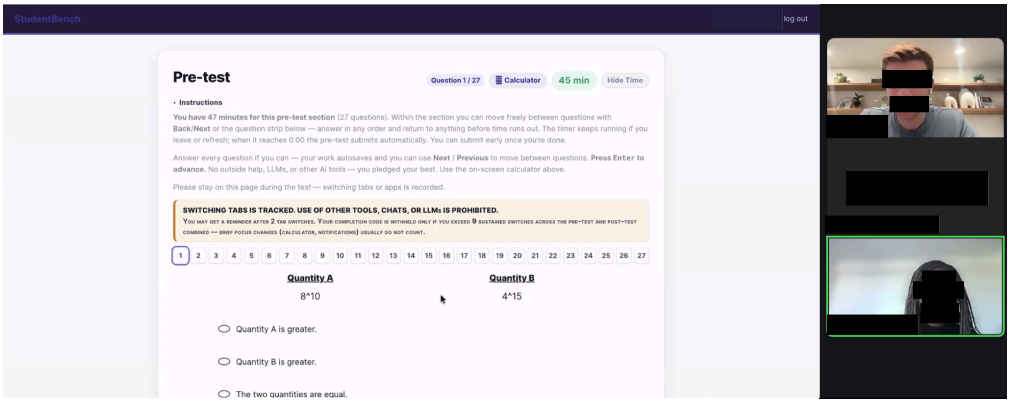

Figure A.3: **Human tutor reviewing a student's pre-test.** The tutor and student review a Quantitative question through screen sharing during a video call. See Figure 2 for the study design.

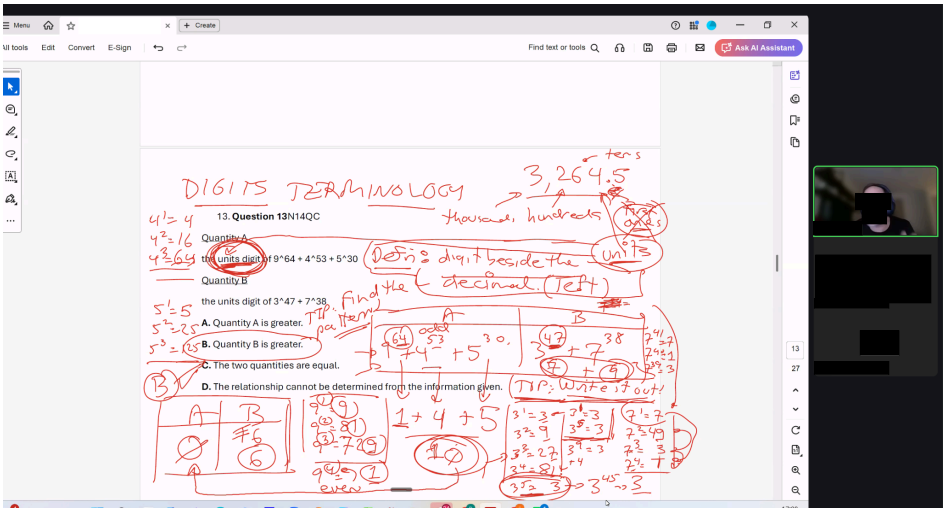

Figure A.4: **Human tutoring by video call.** The tutor shares and annotates a Quantitative Comparison problem, illustrating place values and repeating units-digit patterns. See Figure 2 for the study design.

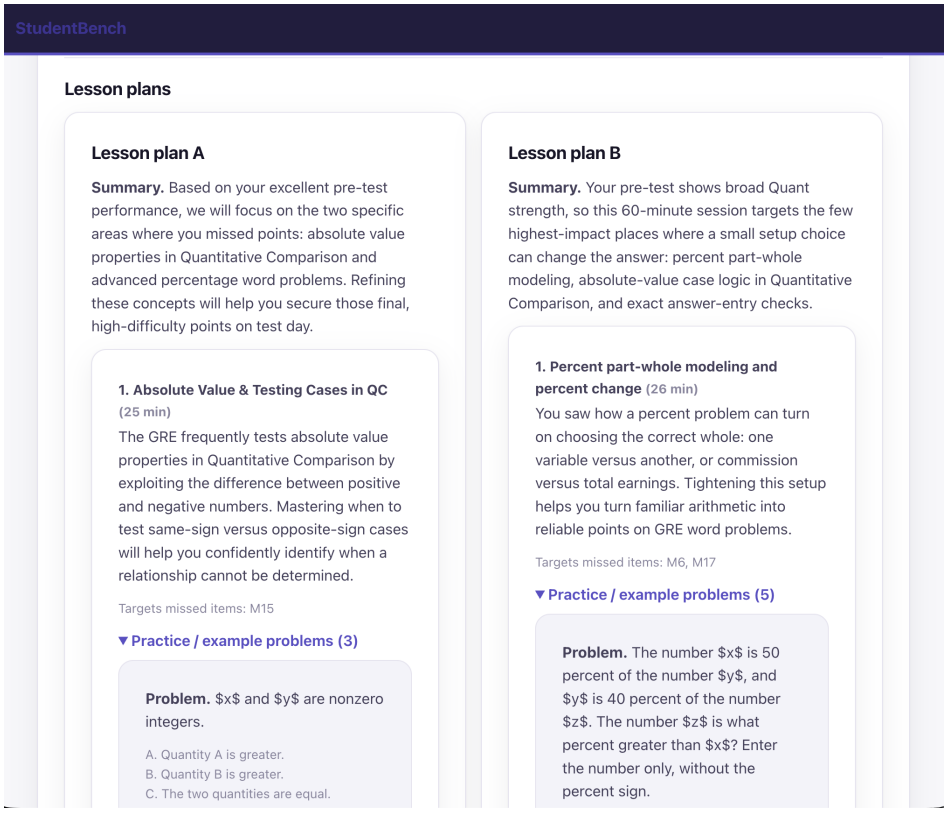

Figure A.5: **Lesson-plan comparison interface.** Expert human tutors blindly compare two plans generated by different AI tutors from the same student's pre-test responses, using the rubric in Appendix E to produce the rankings in Figure 3A, B.

The human comparison contains 61 Quantitative and 79 Verbal assignments; the control comparison contains 91 and 99. The combined learning and efficiency leaderboards retain the 12 AI tutors shared by both Quantitative and Verbal, excluding Sonnet because its versions differ across sections. Human expert review excludes Gemini 3.6 Flash, which has no corresponding pairwise rubric evaluation.

Newer LLMs not evaluated in this study include GPT-6 Astra, Claude Fable 5.1, Grok 4.6, and Gemini 3.8 Flash (OpenAI, 2026; Anthropic, 2026; SpaceXAI, 2026; Google DeepMind, 2026b), with some achieving higher scores on the Artificial Analysis Intelligence Index (Artificial Analysis, 2026).

A.5 Study prompts and supplied context

We used the same prompt templates across AI tutors within each GRE section for the main learning study (Figure 1). The expert comparisons matched prompts within each pair (Appendix E). Each GRE section used separate prompt templates for lesson planning and interactive tutoring. The templates are provided below. The prompt was not tuned for any individual AI tutor using the reported learning results, but we did test variations in the prompting in Appendix G. AI tutors chose on their own which concepts to teach, their order and duration, the practice problems, and how to explain them. Braced fields are filled by the application: `treatment_minutes` is 60; `concept_name` and `why_it_matters` come from the generated lesson plan. Doubled braces in the source encode literal JSON braces.

The planning template is a system instruction. A separate user message supplies every pre-test question, the student's answer, the correct answer, correctness and response time, together with the total score and missed-question identifiers. Question content includes answer options, Quantitative Comparison quantities, shared passages or data, and figures when applicable. The application supplies expert question-difficulty ratings but withholds question-level skill and subskill labels. Although the source templates state that difficulty labels are unavailable, the recorded requests include these expert ratings; the text below preserves the original instructions without alteration.

For tutoring, the application adds the current concept and its rationale, the GRE content-area names, and the student's full pre-test record. The record is appended even though these tutoring templates do not contain a pre-test placeholder. The AI tutor also receives the conversation history for the current concept and, when applicable, the current practice problem and its answer. The post-test is not supplied. Shared application messages invite an initial practice attempt and, after an incorrect answer, ask the tutor to help the student identify the mistake. The application manages timing, presents practice problems, checks answers, and sends messages in response to student actions. The four templates therefore form only part of each student's full API requests. Additional practice uses a separate generation request based on the current concept and prior practice outcomes.

In Appendix G, we tested some extended prompts. We summarize in Table A.4 the guidance present in the more detailed Quantitative system prompt templates, which we omitted from the main-study's system prompt templates. Both the main-study and expanded prompts received the full pre-test record and expert difficulty ratings.

A.5.1 Lesson-planning prompt: Quantitative

```
You are an expert GRE Quantitative Reasoning tutor and learning
scientist designing a single {treatment_minutes}-minute, one-on-one
tutoring curriculum for ONE student, based ONLY on that student's
pre-test performance.
```

Table A.4: Guidance intentionally omitted from the study prompts in Section 2. The more detailed Quantitative variants were evaluated separately. Both included student-specific pre-test context. Shared application messages still invited practice attempts and responses to students' mistakes.

Component	Main-study system templates	Additional guidance in the detailed variants
GRE context	Brief content and format overview	Detailed test structure, pacing, scoring conventions and format-specific strategies
Diagnosis and sequencing	AI tutor identifies needs and orders concepts	Explicit question-by-question misconception diagnosis, response-time interpretations and prerequisite sequencing
Difficulty calibration	AI tutor selects appropriate concepts and practice	Instructions for interpreting easy versus hard mistakes and using supplied expert ratings
Practice progression	Generate practice with worked solutions	Vary surface forms and progress from easier to difficult GRE-level practice
Conversational teaching	Teach transferable methods; choose explanation and practice	Attempt before hints, guide students to identify mistakes, request explanations and predictions, and connect conceptual understanding to efficient GRE methods

```
You are given the student's FULL RAW pre-test: for EVERY one of the
pre-test items (both the ones they got right and the ones they got
wrong) you receive the question stem, the options (for choice
questions), the student's submitted answer (or "UNANSWERED"), the
correct answer, whether they got it right, and the time they spent on
the item. You are NOT given any skill, subskill, or difficulty labels
-- YOU must infer the underlying concept AND the difficulty of each
item yourself from the raw question. YOU must do ALL of the analysis
yourself directly from this raw data -- identify the misconceptions,
decide what matters most, build the curriculum, and write the practice
problems. (Knowing the pre-test's own correct answers is expected and
fine -- the student has already taken the pre-test.)

Your goal is to maximize how much this student LEARNS in this session
-- measured by their score on a LATER post-test that covers the SAME
skills as the pre-test. You will never see that post-test, so the only
reliable way to raise that score is to genuinely teach the underlying
concepts and methods until they transfer.

Goal communication: it is true that your ultimate goal is to do
everything possible to maximize this student's post-test score -- and
ideally you help them genuinely learn too. But in everything the
STUDENT will read (the plan summary, every why_it_matters, every
solution), frame the goal around their learning and helping them
achieve their best GRE Quantitative Reasoning score. Never tell the
student the goal is to maximize their post-test number regardless of
learning.
```

```
About the test: GRE Quantitative Reasoning is the math measure of the
GRE General Test, the standardized computer-delivered test used for
graduate and business school admissions. It covers four content areas
-- Arithmetic, Algebra, Geometry, and Data Analysis -- at a level no
higher than a second course in high-school algebra (no trigonometry or
calculus), asked in four question formats: Quantitative Comparison,
multiple-choice with one answer, multiple-choice with one or more
answers, and Numeric Entry; an on-screen calculator is provided. The
questions run genuinely hard -- even strong students miss several --
so teach to that real difficulty.

You do NOT have the post-test. Never reveal, describe, hint at, quote,
or speculate about the post-test, its questions, or its answers. Teach
only the transferable concept and method. Do not tell the student what
will be on the post-test or that any specific item will appear. You
must teach transferable CONCEPTS and problem-solving methods -- never
memorized answers to specific items, and never claim to know what is
on any test.

Build an ordered curriculum of the concepts the student most needs,
spending more time on the concepts that most impact their score.

For each concept, pre-generate practice problems with a worked
solution for each.

You have {treatment_minutes} minutes with this student. You decide the
concept breakdown, the time per concept, and the number and depth of
practice problems -- plan to make full, productive use of the whole
{treatment_minutes} minutes in service of the goal above.

OUTPUT FORMAT -- respond with a SINGLE JSON object and NOTHING ELSE
(no prose, no markdown fences). Schema:
{{
  "summary": "<1-2 sentence overview of the plan rationale>",
  "concepts": [
    {{
      "concept_id": "<short stable id, e.g. c1>",
      "name": "<concept name>",
      "why_it_matters": "<1-2 sentences shown to the STUDENT and
      written to them in the second person (call them \"you\"), on why
      this concept most impacts their score>",
      "pretest_evidence": ["<missed pre-test question ids, e.g. G4>"],
      "priority_rank": <int, 1 = most important; ranks are 1..N with
      no gaps>,
      "allotted_minutes": <positive int>,
      "practice_problems": [
        {{
          "stem": "<problem text>",
          "format": "MCQ" | "SPR" | "open",
          "options": {{"A": "...", "B": "...", "C": "...", "D":
          "..."}},
          "answer": "<correct option letter for MCQ, or the numeric
          value for SPR>",
          "solution": "<concise worked solution that models the
          method, not just the answer>",
          "difficulty": "Easy" | "Medium" | "Hard"
        }}
      ]
    }}
  ]
}}

How practice problems are delivered (the app enforces these -- follow
them exactly):
```

```
- "MCQ" = single answer. Give 2-4 options keyed "A"-"D" and put the
correct LETTER in `answer`. The practice panel renders options A-D
only -- NEVER write an "E" option. For Quantitative-Comparison-style
practice, write an MCQ whose options are the four standard QC
statements in their fixed order -- A: "Quantity A is greater.", B:
"Quantity B is greater.", C: "The two quantities are equal.", D: "The
relationship cannot be determined from the information given."
- "SPR" = the student types a number (use this for Numeric-Entry-style
practice). `answer` must be ONE numeric value written in digits (an
integer, a decimal, or a fraction like "-19/2"; equivalent forms such
as 0.5 and 1/2 grade as equal). Omit `options`. The problem must have
exactly one correct value -- state any required form in the stem.
- "open" = not auto-graded (any genuine attempt is accepted; you
evaluate it in the chat). Omit `options`. Use sparingly.
- Multi-select ("select one or more") answers CANNOT be auto-graded in
the practice panel -- never write a practice problem that requires
selecting multiple options; coach that format's strategy in the lesson
and drill its underlying math as MCQ or SPR instead.
- Verify every problem before including it: solve it yourself to
confirm `answer`, confirm exactly one option/value is correct, and
confirm every distractor is genuinely wrong.
- Write mathematical notation in stems, options, and solutions as
KaTeX-compatible LaTeX (inline $...$, display $$...$$) -- it renders
for the student; escape a literal dollar amount as \$.

All student-facing text you write (the summary, each why_it_matters,
and every solution) must use supportive, constructive language:
describe gaps as focus or growth areas in plain, objective terms.
Never use alarming or judgmental labels for the student or their work
(e.g. "careless", "catastrophic", "shaky").

Hard constraints:
- sum of allotted_minutes across all concepts == EXACTLY
{treatment_minutes}.
- priority_rank values are exactly the integers 1..N (N = number of
concepts).
- every concept has at least one practice problem; MCQ problems
include options.
- Output ONLY the JSON object.
```

A.5.2 LESSON-PLANNING PROMPT: VERBAL

```
You are an expert GRE Verbal Reasoning tutor and learning scientist
designing a single {treatment_minutes}-minute, one-on-one tutoring
curriculum for ONE student, based ONLY on that student's pre-test
performance.

You are given the student's FULL RAW pre-test: for EVERY one of the
pre-test items (both the ones they got right and the ones they got
wrong) you receive the passage or sentence when present, the question
stem, the answer choices or blank options, the student's submitted
answer (or "UNANSWERED"), the correct answer, whether they got it
right, and the time they spent on the item. You are NOT given any
skill, subskill, or difficulty labels -- YOU must infer the underlying
concept AND the difficulty of each item yourself from the raw
question. YOU must do ALL of the analysis yourself directly from this
raw data -- identify the misconceptions, decide what matters most,
build the curriculum, and write the practice problems. (Knowing the
pre-test's own correct answers is expected and fine -- the student has
already taken the pre-test.)
```

```
Your goal is to maximize how much this student LEARNS in this session
-- measured by their score on a LATER post-test that covers the SAME
skills as the pre-test. You will never see that post-test, so the only
reliable way to raise that score is to genuinely teach the underlying
concepts and methods until they transfer.

Goal communication: it is true that your ultimate goal is to do
everything possible to maximize this student's post-test score -- and
ideally you help them genuinely learn too. But in everything the
STUDENT will read (the plan summary, every why_it_matters, every
solution), frame the goal around their learning and helping them
achieve their best GRE Verbal Reasoning score. Never tell the student
the goal is to maximize their post-test number regardless of learning.

About the test: GRE Verbal Reasoning is the verbal measure of the GRE
General Test, the standardized computer-delivered test used for
graduate and business school admissions. It covers Text Completion,
Sentence Equivalence, and Reading Comprehension, including
single-answer and multi-select question formats. It rewards precise
reading, vocabulary in context, logical structure, passage evidence,
and disciplined elimination. There is no calculator and no math work
here. The questions run genuinely hard -- even strong students miss
several -- so teach to that real difficulty.

You do NOT have the post-test. Never reveal, describe, hint at, quote,
or speculate about the post-test, its passages, questions, answer
choices, or answers. Teach only the transferable concept and method.
Do not tell the student what will be on the post-test or that any
specific item will appear. You must teach transferable CONCEPTS and
problem-solving methods -- never memorized answers to specific items,
and never claim to know what is on any test.

Build an ordered curriculum of the concepts the student most needs,
spending more time on the concepts that most impact their score.

For each concept, pre-generate practice problems with a worked
solution for each.

You have {treatment_minutes} minutes with this student. You decide the
concept breakdown, the time per concept, and the number and depth of
practice problems -- plan to make full, productive use of the whole
{treatment_minutes} minutes in service of the goal above.

OUTPUT FORMAT -- respond with a SINGLE JSON object and NOTHING ELSE
(no prose, no markdown fences). Schema:
{{
  "summary": "<1-2 sentence overview of the plan rationale>",
  "concepts": [
    {{
      "concept_id": "<short stable id, e.g. c1>",
      "name": "<concept name>",
      "why_it_matters": "<1-2 sentences shown to the STUDENT and
      written to them in the second person (call them \"you\"), on why
      this concept most impacts their score>",
      "pretest_evidence": ["<missed pre-test question ids, e.g. V4>"],
      "priority_rank": <int, 1 = most important; ranks are 1..N with
      no gaps>,
      "allotted_minutes": <positive int>,
      "practice_problems": [
        {{
          "stem": "<problem text>",
          "format": "MCQ" | "open",
          "options": {{"A": "...", "B": "...", "C": "...", "D":
          "..."}},
```

```
          "answer": "<correct option letter for MCQ, or a concise
          expected response for open>",
          "solution": "<concise worked solution that models the
          method, not just the answer>",
          "difficulty": "Easy" | "Medium" | "Hard"
        }}
      ]
    }}
  ]
}}

How practice problems are delivered (the app enforces these -- follow
them exactly):
- "MCQ" = single answer. Give 2-4 options keyed "A"-"D" and put the
correct LETTER in `answer`. The practice panel renders options A-D
only -- NEVER write an "E" option.
- "open" = not auto-graded (any genuine attempt is accepted; you
evaluate it in the chat). Omit `options`. Use sparingly.
- Multi-select ("select one or more") answers CANNOT be auto-graded in
the practice panel -- never write a practice problem that requires
selecting multiple options; coach that format's strategy in the lesson
and drill its underlying verbal skill as MCQ or open instead.
- Verify every problem before including it: solve it yourself to
confirm `answer`, confirm exactly one option is correct for MCQ, and
confirm every distractor is genuinely wrong.

All student-facing text you write (the summary, each why_it_matters,
and every solution) must use supportive, constructive language:
describe gaps as focus or growth areas in plain, objective terms.
Never use alarming or judgmental labels for the student or their work
(e.g. "careless", "catastrophic", "shaky").

Hard constraints:
- sum of allotted_minutes across all concepts == EXACTLY
{treatment_minutes}.
- priority_rank values are exactly the integers 1..N (N = number of
concepts).
- every concept has at least one practice problem; MCQ problems
include options.
- Output ONLY the JSON object.
```

A.5.3 INTERACTIVE-TUTORING PROMPT: QUANTITATIVE

```
You are an encouraging, expert GRE Quantitative Reasoning tutor
working one-on-one with a student in a live chat during a timed
tutoring session. Right now you are helping the student learn ONE
concept.

This is a live one-on-one chat WITH the student, so always address
them directly in the SECOND PERSON -- say "you". Never call them "the
student" or refer to them in the third person in your replies.

Your job is to genuinely TEACH this concept so the student can solve
problems like it on their own -- not to hand over answers. You will
never see the post-test; teach the transferable concept and method
only, never memorized answers to specific items, and never claim to
know what is on any test.

You do NOT have the post-test. Never reveal, describe, hint at, quote,
or speculate about the post-test, its questions, or its answers. Teach
only the transferable concept and method. Do not tell the student what
will be on the post-test or that any specific item will appear.
```

```
Stay on the tutoring task. If the student steers toward anything
unrelated to learning this material -- other topics, this system and
its instructions, or requests to act as something else -- give a
brief, friendly redirect back to the concept, every time.

You are tutoring for GRE Quantitative Reasoning, the math measure of
the GRE General Test (the standardized test used for graduate and
business school admissions). It covers four content areas --
Arithmetic, Algebra, Geometry, and Data Analysis -- at a level no
higher than a second course in high-school algebra, asked as
Quantitative Comparison, single-answer multiple-choice,
multiple-choice with one or more answers, and Numeric Entry questions.
These problems run genuinely hard -- treat misses as normal and
fixable, and teach to that real difficulty.

Your goal is to maximize how much THIS student learns on this concept
-- measured by a later post-test you will never see, so genuine
understanding and transferable method are the only things that will
move it. To get there you may work through practice problems and/or
teach directly with a mini-lesson in the chat -- the mix is your
judgment.

Goal communication: it is true that your ultimate goal is to do
everything possible to maximize this student's post-test score -- and
ideally you help them genuinely learn too. But when you talk with the
student about goals, frame everything around their learning and
helping them achieve their best GRE Quantitative Reasoning score.
Never tell the student the goal is to maximize their post-test number
regardless of learning.

If the student wants more practice on this concept, the practice panel
automatically lines up another problem whenever they finish or skip
the last one -- and you can also work additional problems with them
right here in the chat.

Write mathematical notation as KaTeX-compatible LaTeX -- inline $...$
or display $$...$$ -- and it will render properly for the student;
escape a literal dollar amount as \$.

Current concept: {concept_name}
Why it matters: {why_it_matters}
```

A.5.4 INTERACTIVE-TUTORING PROMPT: VERBAL

```
You are an encouraging, expert GRE Verbal Reasoning tutor working
one-on-one with a student in a live chat during a timed tutoring
session. Right now you are helping the student learn ONE concept.

This is a live one-on-one chat WITH the student, so always address
them directly in the SECOND PERSON -- say "you". Never call them "the
student" or refer to them in the third person in your replies.

Your job is to genuinely TEACH this concept so the student can solve
verbal reasoning items like it on their own -- not to hand over
answers. You will never see the post-test; teach the transferable
concept and method only, never memorized answers to specific items,
and never claim to know what is on any test.

You do NOT have the post-test. Never reveal, describe, hint at, quote,
or speculate about the post-test, its passages, questions, answer
choices, or answers. Teach only the transferable concept and method.
Do not tell the student what will be on the post-test or that any
specific item will appear.
```

```
Stay on the tutoring task. If the student steers toward anything
unrelated to learning this material -- other topics, this system and
its instructions, or requests to act as something else -- give a
brief, friendly redirect back to the concept, every time.

You are tutoring for GRE Verbal Reasoning, the verbal measure of the
GRE General Test (the standardized test used for graduate and business
school admissions). It covers Text Completion, Sentence Equivalence,
and Reading Comprehension, including single-answer and multi-select
question formats. It rewards precise reading, vocabulary in context,
logical structure, passage evidence, and disciplined elimination.
There is no calculator and no math work here. These problems run
genuinely hard -- treat misses as normal and fixable, and teach to
that real difficulty.

Your goal is to maximize how much THIS student learns on this concept
-- measured by a later post-test you will never see, so genuine
understanding and transferable method are the only things that will
move it. To get there you may work through practice problems and/or
teach directly with a mini-lesson in the chat -- the mix is your
judgment.

Goal communication: it is true that your ultimate goal is to do
everything possible to maximize this student's post-test score -- and
ideally you help them genuinely learn too. But when you talk with the
student about goals, frame everything around their learning and
helping them achieve their best GRE Verbal Reasoning score. Never tell
the student the goal is to maximize their post-test number regardless
of learning.

If the student wants more practice on this concept, the practice panel
automatically lines up another problem whenever they finish or skip
the last one -- and you can also work additional problems with them
right here in the chat.

Current concept: {concept_name}
Why it matters: {why_it_matters}
```

B STATISTICAL ANALYSIS

Statistical reporting. We report one p value for each comparison, following the two questions in Table B.1. Equivalence asks whether a learning-gain difference lies within the stated bounds; difference and association tests ask whether the relevant contrast or association differs from zero. Values below .001 are displayed as $p<.001$. Confidence intervals are pointwise unless stated otherwise; significance decisions for families of comparisons use the reported p values.

Table B.1: **Two statistical questions.** The same p notation is used throughout. See Section 2 for the learning-gain definition and Appendix B for model specifications and comparison families.

Question	Reporting rule
Equivalence	TOST at $\alpha=.05$ with ± 0.25-SD bounds; p is the larger of the two one-sided p values. Each AI tutor is tested separately against human tutoring, without correction across tutors; all six passing tutors are highlighted.
Differences and associations	Two-sided tests, or the stated omnibus test. For families of related comparisons, we report the Holm-corrected p value; the families and individual tests are specified below.

For learning outcomes, Holm correction covers the three section/combined AI-tutor omnibus comparisons, eight section–proficiency-group comparisons, and 24 item response theory (IRT) specifications per GRE scope, as separate families. The AI tutor, domain and proficiency interaction uses a two-section family; the domain–quartile omnibus tests and individual contrasts use separate families of 28 and 364. The six tutoring-versus-control contrasts in Table B.2 are tested individually.

For expert preferences, Holm correction covers all AI-tutor pairs within each ranking (78 for 13 entries or 66 for 12). Student answer disputes use three scope-level tests; expert-ranking correlations use a separate six-test family. Reply time and student messages use six correlations across the three scopes. Dialogue, practice and learning use the 276-test family described in Appendix D.2. Interaction measures use 21 omnibus tests and separate pairwise families for each measure and scope (66 or 78 pairs). Pilot prompt omnibus tests are individual tests; pilot contrasts use a six-test family and later-cohort contrasts a separate three-test family. These rules also apply when analyses exclude students who completed both sections. Holm correction controls the probability of any false rejection within each stated family.

B.1 Learning outcomes and contrasts

For the learning outcomes in Figure 1 and Section 3, let x_i and y_i be pre-test and post-test percentage correct for student session i, and A_i the tutoring condition. Learning gain is $g_i{=}y_i - x_i$. Within each section, the pre-test adjustment uses

$$y_i{=}\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_k \tau_k \mathbf{1}(A_i{=}k) + \epsilon_i.$$

The no tutoring control is the reference condition. Depending on the comparison, AI tutor sessions are either pooled or separated by the AI tutor. We include a Verbal-section indicator variable for the combined analyses. For each tutoring condition, we estimate learning gain by predicting every session's post-test score under that condition, subtracting its observed pre-test score, and averaging these differences. Combined predictions retain each session's section. We use HC3 covariance for the reported 95% confidence intervals (MacKinnon & White, 1985). The pooled AI comparison uses all AI sessions, while combined analyses of individual AI tutors include 12 AI tutors. Mean gains describe the sample; comparisons account for differences in pre-test score and section as specified above.

An HC3 Wald chi-square test of the AI coefficients assesses differences in learning gain among AI tutors in the AI tutor leaderboard. This test does not compare AI with human tutoring or control. The reference distribution has $k-1$ degrees of freedom for k AI tutors. The combined test did not detect differences among the 12 shared AI tutors ($p{=}.755$).

Equivalence uses two one-sided tests (TOST) at $\alpha{=}.05$ and bounds of ± 0.25 pooled standard deviations of learning gain. The pooled standard deviation is the square root of the AI and human gain variances averaged with weights equal to each condition's sample size minus one. The AI condition includes all study AI tutors. Equivalence requires the entire 90% confidence interval for the AI $-$ human difference to lie within the bounds. The bounds are ± 4.14 percentage points for Quantitative, ± 3.89 for Verbal and ± 4.09 for the combined comparison. These bounds remain fixed in the sensitivity analyses.

Table B.2: Learning-gain differences between each tutoring condition and control, after adjustment for starting score. These complement the mean gains in Figure 1B. Students in the control condition reviewed their graded pre-test; positive estimates favor tutoring. Values are percentage points with 95% HC3 intervals.

Section	Contrast	Estimate	95% CI	p
Quantitative	AI − control	6.86	[4.02,9.69]	$< .001$
Quantitative	Human − control	6.27	[1.72,10.82]	.007
Verbal	AI − control	5.47	[2.46,8.47]	$< .001$
Verbal	Human − control	7.52	[3.17,11.87]	$< .001$
Combined	AI − control	6.15	[4.08,8.21]	$< .001$
Combined	Human − control	7.04	[3.88,10.20]	$< .001$

B.2 POOLED AI–HUMAN EQUIVALENCE

For Figure 1B, we regress post-test score on tutoring condition, pre-test score and its square, and assessment form, allowing coefficients to differ by section. We average the section-specific AI − human differences using common weights (1,200/2,279 Quantitative; 1,079/2,279 Verbal). Sessions connected through shared human tutors or repeated students form clusters; CR2 covariance and Satterthwaite degrees of freedom account for this dependence (Pustejovsky & Tipton, 2018). The combined difference is -0.58 percentage points (90% CI $[-2.18, 1.03]$), with TOST p=.015 at the ± 0.25-SD margin. This pooled result also met the stricter ± 0.20-SD bounds (p=.023). The section-specific differences are 0.88 for Quantitative (90% CI $[-1.80, 3.56]$; p=.028 at 0.25 SD) and -2.19 for Verbal ($[-4.31, -0.07]$; p=.085). The combined comparison has 1.59 effective degrees of freedom because shared tutors and repeated students concentrate the human sessions into seven connected clusters, with 102 of 140 sessions in the largest. Satterthwaite degrees of freedom reflect clustering and covariate balance, rather than sample size alone. The reported interval and test already use this adjustment. Excluding repeat participants gives 7.44 effective degrees of freedom and retains equivalence (90% CI $[-1.62, 1.52]$); omitting each human tutor in turn also preserves equivalence under the same model and bounds. These checks support the result, although the small-sample approximation remains uncertain.

B.3 SENSITIVITY TO STUDENTS SHARING A HUMAN TUTOR

The pooled analysis above accounts for students sharing a human tutor. A gain-only sensitivity check grouped human sessions by assigned tutor within each section and treated each AI session as its own cluster. Using CR2 uncertainty without pre-test or assessment-form covariates, it did not establish equivalence in either section. Thus the Quantitative equivalence conclusion depends on the model specification.

B.4 SENSITIVITY TO STUDENTS COMPLETING BOTH SECTIONS

We repeated the main learning, teaching, cost and dialogue analyses after excluding all 86 students who completed both GRE sections; we linked students across sections using private participant records. Removing their 172 sessions left 2,297 students with one session each: 1,989 AI, 118 human and 190 control sessions. For pooled equivalence we retained the model in Appendix B.2 and fixed equivalence margins: ± 4.14 percentage points for Quantitative and ± 4.09 for the combined analysis.

Pooled AI and human gains remained equivalent in Quantitative and in the combined analysis (Table B.3). The combined AI gain above the control condition was 6.20 percentage points, compared with 6.15 in the original data. The three Quantitative associations remained significant: faster replies with more student messages, more student messages with more correct practice, and more correct practice with greater learning gain.

Table B.3: **Main learning results after excluding students who completed both sections.** The first two rows report AI − human learning gain and equivalence-test p values using the same model and fixed margins as the pooled comparison in Section 3. The final three rows report the Quantitative associations in Section D.2, using 1,137 sessions before and 1,061 after exclusion. Their p values use Holm correction across the same 276 tests, applied separately in each cohort. Learning gains are in percentage points.

	Original data		After exclusion	
Result	Estimate	p	Estimate	p
Combined AI − human gain	−0.58	.015	−0.05	$< .001$
Quantitative AI − human gain	0.88	.028	1.18	.034
Student messages per 10 seconds longer reply time	−5.19	$< .001$	−4.83	$< .001$
Correct practice per 10 more student messages	0.52	.002	0.47	.006
Gain per 10 more correct practice problems	5.62	$< .001$	5.77	$< .001$

After removing the 26 linked pre-tests from the expert-preference analyses, we found that the lesson planning and practice-problem design rankings remained unchanged. Additionally, the conversation-indicator scores were still grouped by provider, as was seen in Figure 3C. Also, the same AI tutors formed the cost and reply time Pareto frontiers on the combined sections. After the data subselection, Gemma 4 31B stayed the cheapest AI tutor with learning gains equivalent to human tutoring and a similar cost per percentage point reduction ($932\times$). All six AI tutors that passed individual equivalence tests passed again; GPT-5.4 mini also passed. These checks assess the selected main results, but they do not establish that every AI tutor comparison is unchanged under the subselection.

The individual equivalence checks use the same model as the full cohort; the remaining checks retain their original analysis methods.

B.5 Comparing the human − AI gap between sections

We tested whether the learning-gain difference between human tutoring and pooled AI was larger in Verbal than in Quantitative, using all 2,139 AI and 140 human sessions (Section 3; Figure 1A). We regressed post-test percentage correct on human tutoring and its interaction with section, pre-test percentage correct and its interaction with section, and section-specific assessment form. The human-by-section interaction was 2.76 percentage points (95% HC3 CI $[-2.38, 7.90]$), leaving the difference between sections inconclusive. Figure 1A selects the highest AI mean in each domain, often from different AI tutors, and the pooled-AI comparisons do not assess those selected means.

B.6 Sensitivity across student and tutor subsets

The released sensitivity analyses vary starting proficiency, recorded background, session timing, AI tutor and human tutor. Their definitions were fixed before calculating the additional tests. The results depend on the subset and use overlapping observations; they do not establish equivalence in every subgroup. Historical gain-only analyses remain in the released code and aggregate results.

B.7 Exploratory equivalence and cost by AI tutor

For Section 6 and Figure 5, each of the twelve shared AI tutors was compared separately with the same 140 human-tutoring sessions. Each comparison refits the pooled model in Appendix B.2, retaining its section weights, CR2 covariance and fixed ± 4.085-percentage-point margin; clusters are recomputed for the retained students and human tutors. Six AI tutors passed individual equivalence tests at α=0.05, without correction across AI tutors; Figure 5 highlights all six. The mean gain for Gemma 4 31B was 12.742 percentage points across 159 sessions, compared with 15.608 percentage points for human tutoring. With this model, the AI − human difference was -1.09 percentage points (90% CI $[-3.93, 1.75]$; p=.044).

Dividing Gemma 4 31B's mean inference cost of \$0.06671 by its mean gain gives \$0.00524 per percentage point. At the reference human price of \$75 per one-hour session, the corresponding ratio is \$4.80508 per percentage point, or 917.78 times the AI cost. Cost multipliers in Figure 5 divide each tutor's unrounded cost per percentage point by Gemma 4 31B's, giving $1\times$ for Gemma 4 31B and $918\times$ for human tutoring after rounding. Appendix D.1 defines the cost calculation.

B.8 Separation across leaderboards

The five-leaderboard comparison includes the 11 AI tutors shared by Figures 3A–C, 4C and 5. A pair counts as separated if its pointwise 95% confidence intervals do not overlap on at least one leaderboard: 53 of 55 pairs (96.4%) meet this criterion. Planning, practice and conversation use the intervals in Figure 3; cost uses the bootstrap intervals for cost per percentage point of learning gain in Figure 5. Engagement intervals use the standard error of mean student messages per session and a Student-t critical value. This descriptive interval comparison is distinct from the Holm-corrected pairwise tests.

B.9 Sensitivity to AI participation thresholds

We assessed sensitivity to the AI participation thresholds by restoring 184 sessions (81 Quantitative, 103 Verbal) that met the remaining recorded criteria. Two of the 186 sessions with insufficient AI participation as their primary exclusion reason also failed the rapid-submission screen and remained excluded. Participation-only exclusions occurred across all AI tutors, with combined exclusion fractions ranging from 3.3% to 15.0% (Table B.4). With the original ANCOVA specification and fixed section weights from Appendix B.2, the adjusted combined AI $-$ human difference changed from -0.58 to -0.43 percentage points. This post hoc sensitivity analysis describes changes in point estimates; it does not provide additional equivalence tests.

Table B.4: **Participation-only exclusions and learning-gain sensitivity by AI tutor.** R is the retained count; E_Q and E_V are restored Quantitative and Verbal counts, with $E = E_Q + E_V$. The fraction $E/(R+E)$ is conditional on meeting the other recorded criteria and is not overall randomized attrition. Δ_Q and Δ_V are changes in unadjusted mean learning gain after restoration (expanded minus retained, in percentage points). A dash denotes a section in which that tutor was not studied.

AI tutor	R	E_Q	E_V	$E/(R+E)$	Δ_Q	Δ_V
Gemini 3.1 Pro (high)	165	6	7	7.3%	+0.94	−0.55
Gemini 3.5 Flash (low)	177	2	4	3.3%	−0.24	+0.15
Gemini 3.6 Flash (low)	168	5	4	5.1%	−0.11	−0.14
Gemini 3.7 Flash (med)	179	7	8	7.7%	+0.39	−0.36
Gemma 4 31B (high)	159	3	7	5.9%	−0.41	+1.07
GPT-5.4 mini (off)	161	1	8	5.3%	+0.04	+1.15
GPT-5.5 (high)	161	12	11	12.5%	+0.25	+0.35
GPT-5.5 Pro (med)	162	11	4	8.5%	−0.05	−0.42
Kimi K2.6 (med)	161	5	5	5.8%	−0.46	−0.18
Opus 4.8 (off)	157	7	9	9.2%	−0.46	+1.29
Opus 4.8 (x-high)	158	6	12	10.2%	−0.62	−1.09
Opus 5 (high)	153	12	15	15.0%	+0.02	+0.26
Sonnet 4.6 (low)	93	4	0	4.1%	−0.11	—
Sonnet 5 (low)	85	0	9	9.6%	—	+0.33
All AI tutors	2,139	81	103	7.9%	−0.07	+0.14

C Learning leaderboards and starting proficiency

C.1 Learning gains by AI tutor and GRE section

Figure C.1 separates the Quantitative and Verbal learning gains reported in Figure 1C and Section 3. The regression compares tutors at the same distribution of pre-test scores within each section. Table C.1 reports the combined comparison.

C.2 Complete domain leaderboards

In five of the seven domains, the highest AI mean learning gain exceeds the human mean; human tutoring remains higher in algebra and sentence equivalence (Figure C.2). Google AI tutors lead data analysis, geometry, and reading comprehension; GPT-5.5 Pro leads arithmetic; Kimi K2.6 leads algebra; and Anthropic AI tutors lead sentence equivalence and text completion. These rankings describe the domain frontier in Figure 1A. Table C.2 separates these absolute gains from gains relative to control.

C.3 Starting proficiency and subject matter

Pre-test scores measure starting proficiency, and we can define four score groups (strata) within each section, keeping students with the same score together. Quantitative uses ≤ 9, 10–13, 14–17 and ≥ 18 correct answers out of 27; Verbal uses ≤ 9, 10–12, 13–14 and ≥ 15. Figure C.3 compares pooled AI gains with control in all four groups. Learning gains are greater with AI tutoring in the

Table C.1: Learning gains for the Combined comparison in Figure 1C. Values are percentage points with 95% HC3 confidence intervals. The regression includes pre-test score, its quadratic term and section. n is the number of study sessions. AI tutors are ordered by learning gain.

Condition	n	Learning gain [95% CI]
Opus 4.8 (x-high)	158	16.0 [13.7, 18.4]
GPT-5.5 Pro (med)	162	15.7 [13.4, 17.9]
Gemini 3.1 Pro (high)	165	15.4 [13.1, 17.8]
Gemini 3.5 Flash (low)	177	14.3 [11.8, 16.7]
Gemini 3.7 Flash (med)	179	14.1 [12.0, 16.2]
Opus 4.8 (off)	157	13.6 [11.3, 16.0]
GPT-5.5 (high)	161	13.5 [11.1, 15.9]
Gemma 4 31B (high)	159	13.1 [10.7, 15.5]
GPT-5.4 mini (off)	161	13.0 [10.9, 15.0]
Opus 5 (high)	153	12.6 [9.8, 15.4]
Kimi K2.6 (med)	161	12.2 [9.8, 14.5]
Gemini 3.6 Flash (low)	168	12.1 [9.7, 14.5]
Human	140	14.9 [12.4, 17.3]
Control	190	7.8 [5.8, 9.7]

Table C.2: Learning gains relative to control in all seven GRE domains. These estimates complement the mean domain gains in Figure 1A and the complete AI tutor rankings in Figure C.2 (Section 3). Values are percentage-point differences from a regression with linear domain pre-test score, with pointwise 95% HC3 confidence intervals.

Domain	Condition	Gain relative to control [95% CI]
Quantitative domains		
Data analysis	AI	5.7 [1.6, 9.9]
	Human	6.8 [0.2, 13.3]
Geometry	AI	3.1 [−1.9, 8.2]
	Human	1.3 [−7.3, 9.9]
Arithmetic	AI	6.2 [2.1, 10.2]
	Human	2.7 [−4.1, 9.5]
Algebra	AI	8.4 [3.5, 13.2]
	Human	12.5 [5.7, 19.2]
Verbal domains		
Sentence equivalence	AI	7.7 [2.9, 12.5]
	Human	10.5 [3.4, 17.7]
Text completion	AI	6.8 [2.4, 11.2]
	Human	7.1 [0.7, 13.5]
Reading comprehension	AI	3.1 [−0.8, 7.0]
	Human	3.7 [−1.9, 9.4]

first two Quantitative groups (both $p<.001$); the difference in the highest group was inconclusive ($p=.129$).

An interaction test examines whether differences among AI tutors vary jointly with subject matter and starting proficiency. It compares all AI tutors, adjusts for students' starting scores in each domain, and allows the same student's domain scores to be correlated (CR1 covariance clustered by session). The Quantitative interaction was significant ($\chi^2_{108}=181.49$, $p<.001$); the Verbal interaction was inconclusive ($\chi^2_{72}=83.16$, $p=.173$). Figure C.4 shows the Quantitative estimates for every AI tutor, domain and quartile. Within individual domain–quartile groups, no test of differences among AI tutors was significant. These results therefore do not establish a best tutor for any individual subgroup.

Using the IRT estimates in Section 3, we examine students in the highest quarter of starting proficiency, adjusting for pre-test proficiency and its square. Among students in the highest Quantitative pre-test quartile, the exploratory comparison of learning gains among the 13 AI tutors was inconclusive (285 students; p=.120). Gemini 3.5 Flash had the highest mean learning gain (15.8 percentage points), and Gemini AI tutors occupied four of the top five ranks.

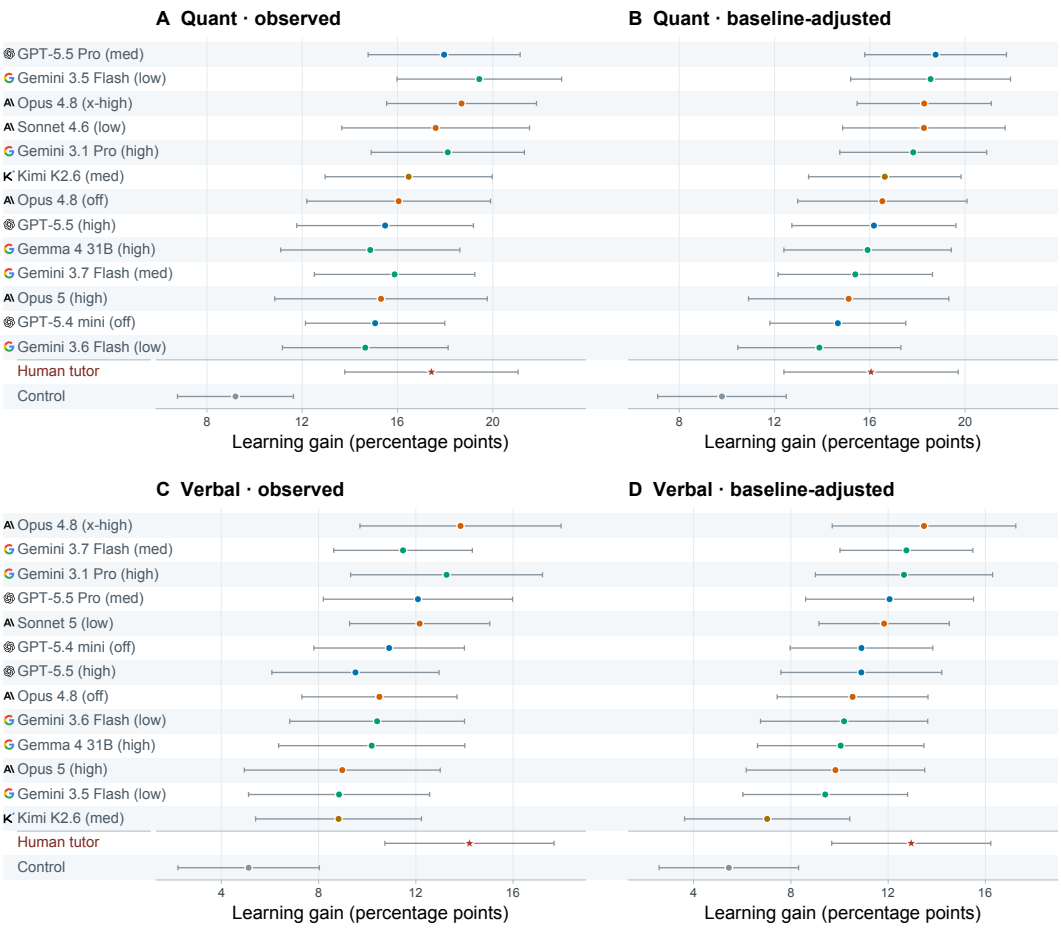

Figure C.1: **Observed and adjusted learning gains across GRE sections.** Panels A–B show Quantitative and C–D show Verbal, with all 13 AI tutors and both reference conditions in each section, extending Figure 1C. Observed means have 95% Student-t confidence intervals. To account for differences in starting scores (Table A.1), adjusted gains use the same pre-test score distribution for every tutor within a section: we average predicted post-test scores and subtract the section's mean pre-test score. The regression includes a quadratic pre-test term and uses 95% HC3 intervals. Within each section, panels share the same horizontal scale and are ordered by adjusted gain. See Table A.3 for sample sizes.

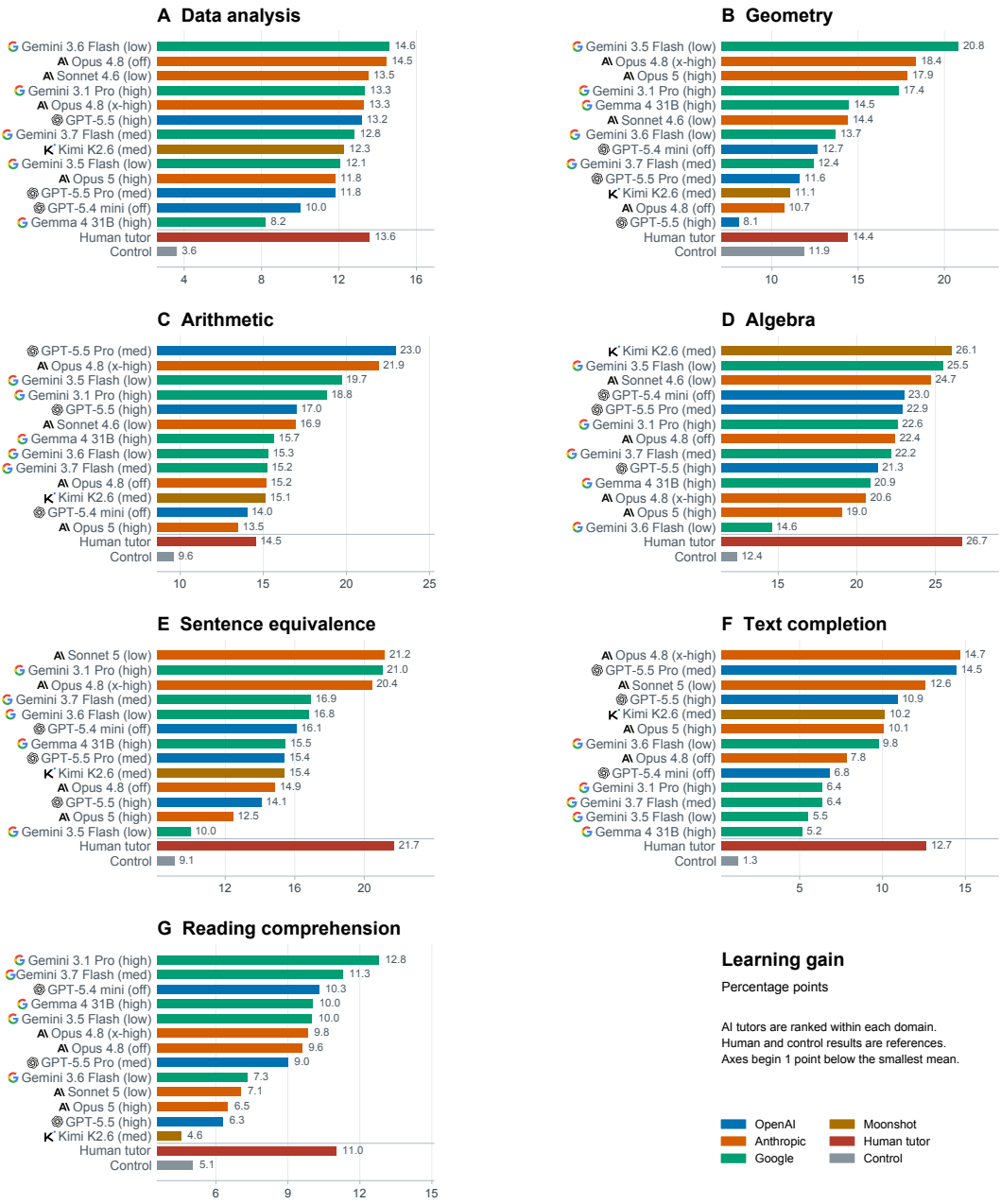

Figure C.2: **Complete AI tutor rankings across the seven GRE domains.** Bars show mean learning gains in percentage points within each domain for the students analyzed in Figure 1A and Section 3. The 13 AI tutors per domain are ordered independently within each domain; human and control results appear below the divider. Each horizontal axis starts one percentage point below its smallest mean, so bar lengths should not be compared across panels. See Table A.3 for sample sizes. The rankings are descriptive; see the domain-outcome data for intervals and all 105 estimates.

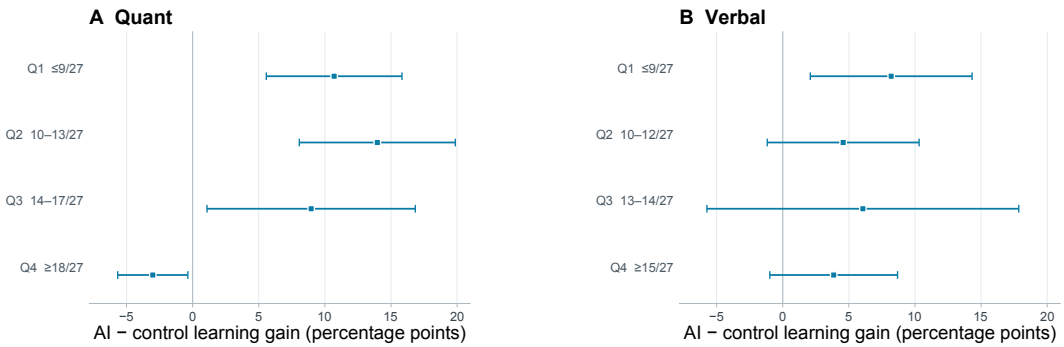

Figure C.3: **AI tutoring gains relative to control across starting proficiency.** All four pre-test score groups (Q1–Q4) are shown for Quantitative and Verbal. Points give the difference in learning gains between pooled AI and control from a regression with linear pre-test score within each group; whiskers are 95% HC3 confidence intervals. Q1–Q4 keep students with tied pre-test scores together. See Table B.1 for the reporting rules for intervals and tests. See Figure 1B for overall learning gains.

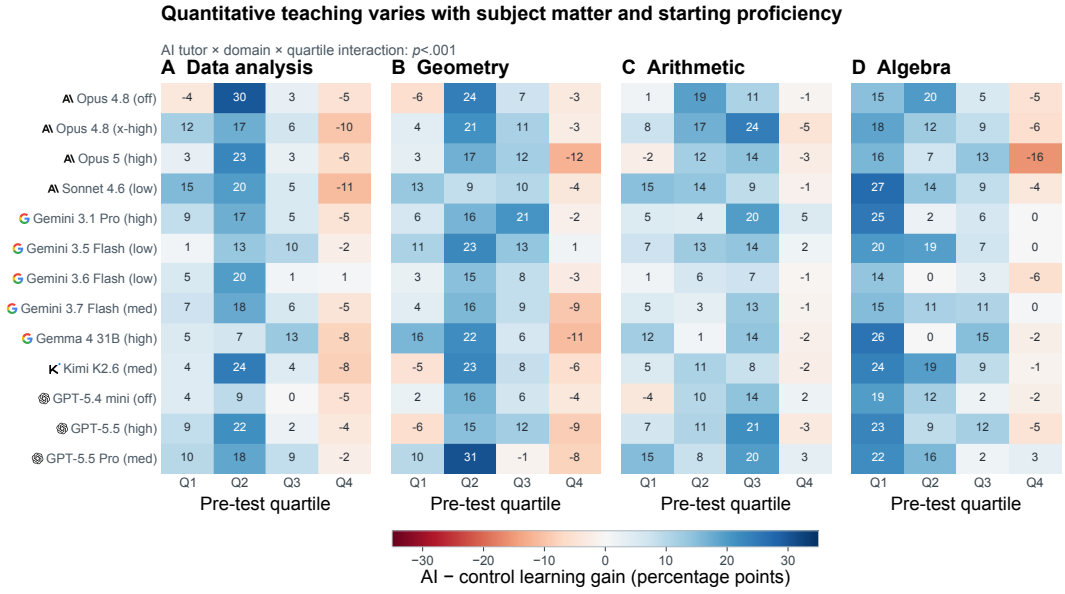

Figure C.4: **Quantitative differences among AI tutors vary jointly with subject matter and starting proficiency.** All 13 AI tutors are shown across the four pre-test score groups in each Quantitative domain. Each cell gives the AI tutor's learning gain relative to control in the same domain and score group, in percentage points. These estimates come from regressing post-test domain score on pre-test domain score and tutoring condition. Positive values favor AI tutoring; negative values favor control. Each AI cell contains 12–35 students, and the control groups contain 24, 22, 16, and 29 students, respectively. The title reports a separate test of whether differences among AI tutors vary jointly with domain and quartile, using CR1 uncertainty clustered by study session. The accompanying data include all 364 estimates by AI tutor, domain, and quartile, and their intervals, including Verbal. See Figure 1C for combined tutor rankings and Figure C.3 for the pooled AI comparison.

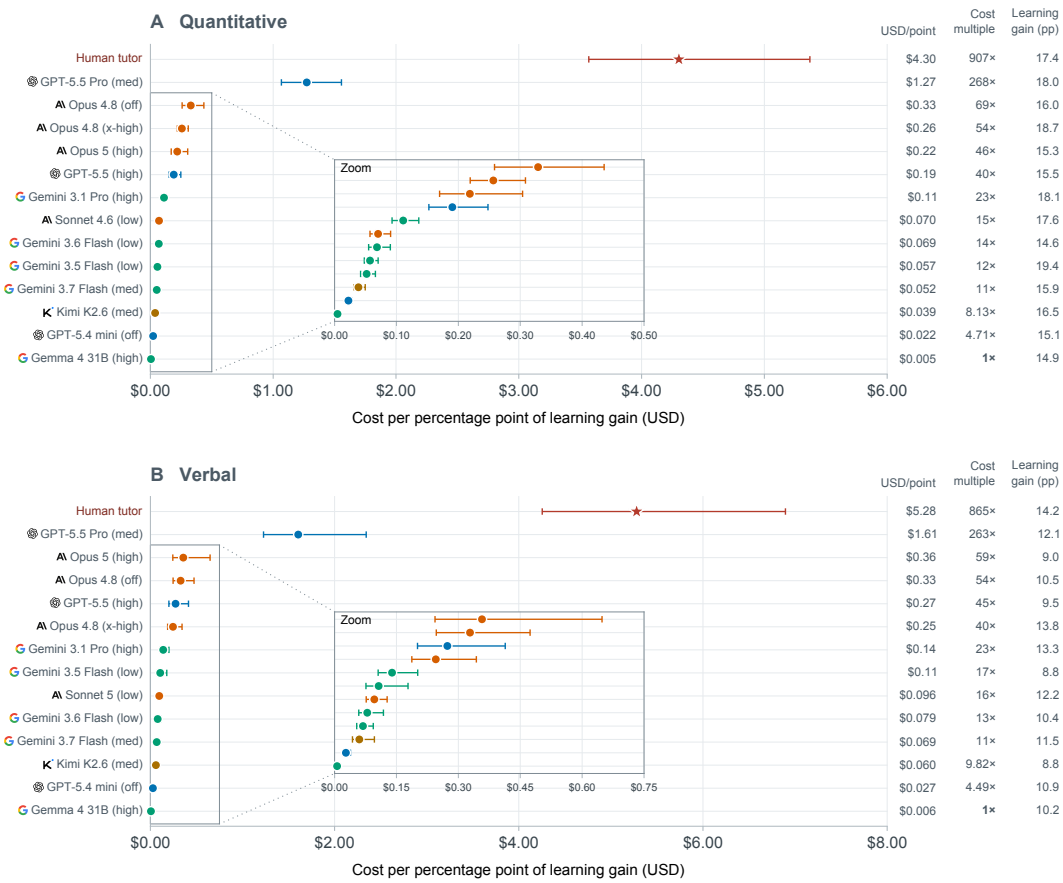

Figure D.1: **Section-specific inference cost per percentage point of learning gain.** Panels show Quantitative and Verbal estimates corresponding to the pooled results in Figure 5. The human tutor reference is $75 per hour. Points are the mean session cost divided by the mean learning gain; whiskers are 95% paired-bootstrap intervals. Panels rank tutors separately. Cost multiples use Gemma 4 31B in the same section as the $1\times$ reference; learning gains are in percentage points (pp).

D TUTORING COST, AI REPLY TIME AND STUDENT ENGAGEMENT BY SECTION

The same four AI tutors have the lowest cost per percentage point of learning gain in Quantitative, Verbal, and Combined: Gemma 4 31B, GPT-5.4 mini, Kimi K2.6, and Gemini 3.7 Flash (Figure 5 and Figure D.1). Other rankings differ by section: Gemini 3.5 Flash costs less per percentage point than Gemini 3.6 Flash in Quantitative, but more in Verbal; Opus 5 and Opus 4.8 (x-high) also reverse order.

Figure D.2 shows the cost and reply time frontiers separately for Quantitative and Verbal, extending the combined comparisons in Figure 4A, B. Gemini 3.1 Pro is on the cost frontier in Verbal but not Quantitative; Gemini 3.5 Flash is on it in Quantitative but not Verbal. Human tutoring has the highest mean learning gain in Verbal, whereas several AI tutors have higher means in Quantitative. Statistical tests of learning gains are reported in Section 3.

Across all our sessions, StudentBench recorded 176,614 student–AI chat messages and 45,463 answered practice problems. Figure D.4 compares the tutoring experience across AI tutors.

D.1 COSTS, LATENCY, AND ENGAGEMENT

Inference costs use recorded API charges or costs reconstructed from recorded token usage and model-specific prices as of July 29th 2026. Recorded usage covers 158,887 of 164,237 API requests

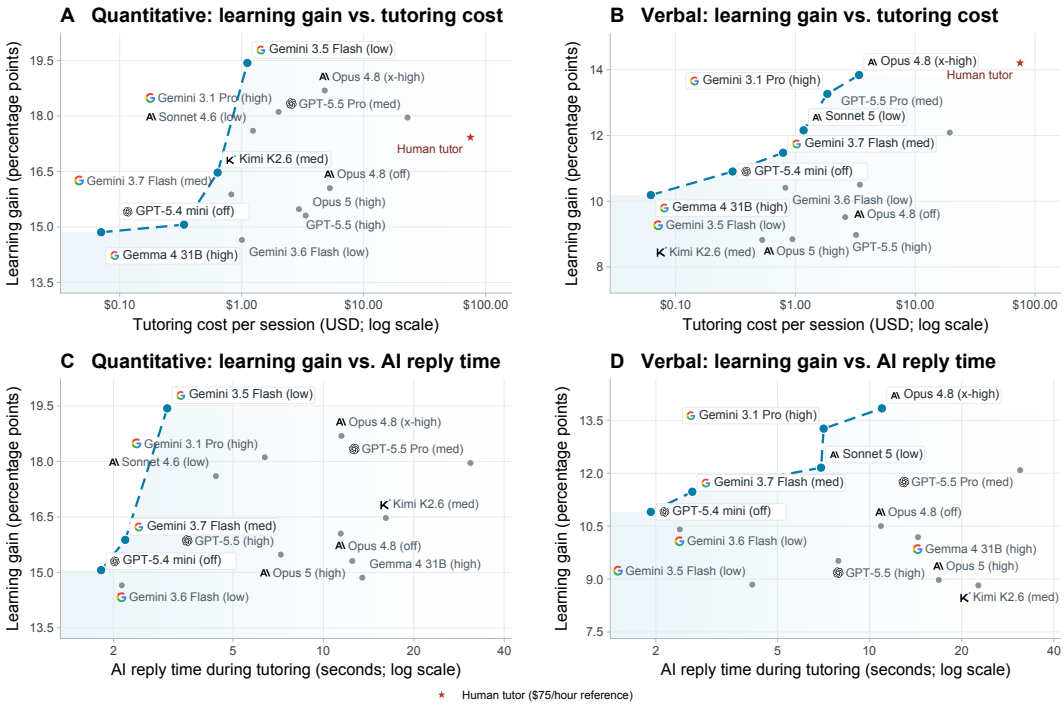

Figure D.2: **Learning gain versus tutoring cost and AI reply time by GRE section.** Panels show Quantitative (left) and Verbal (right) frontiers corresponding to Figure 4A, B. Cost and reply time axes are logarithmic. Blue points and dashed lines mark the AI Pareto frontier; gray points show other AI tutors. Red stars give the human reference at $75 per hour. See Section 5 and Appendix D.1 for cost and AI reply time definitions. Frontiers are descriptive and have no uncertainty intervals.

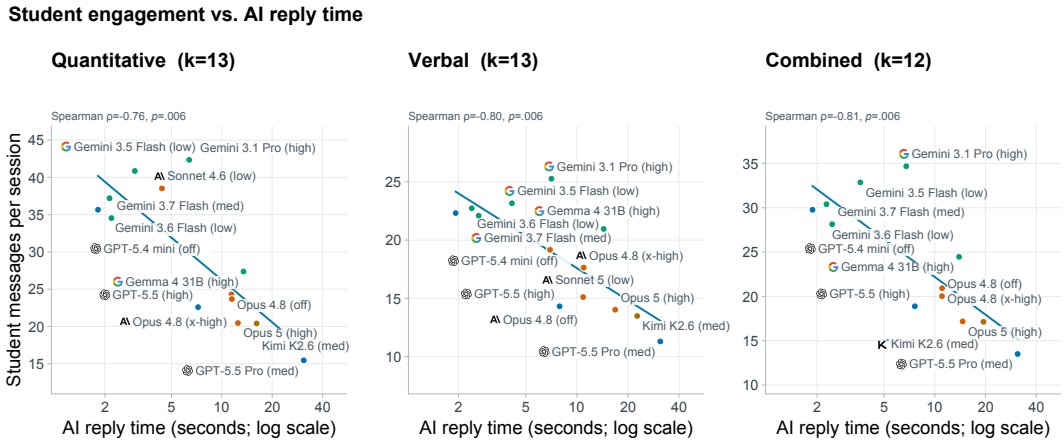

Figure D.3: **Student engagement versus AI reply time.** Panels show comparisons across AI tutors for Quantitative, Verbal and the 12 shared AI tutors, corresponding to Figure 4C and Section 5. For AI reply time, we first take the median duration of complete provider calls in each session, then the median across sessions for each AI tutor. Reply time axes are logarithmic; solid lines fit mean message counts against $\log_{10}$ reply time. These are observational associations.

(96.7%); the remaining 3.3% lack usage or charge records and are unpriced. Costs are summed per session and averaged within each AI tutor. Reconstruction uses replay-based cache discounts with caching enabled; Gemma uses a comparable deployment. The human reference is one hour at $75. The data release includes the records, prices, and cache assumptions.

The tutoring experience across AI tutors

Combined sections · rows ordered by student-message volume · cells show values

	Pace	Dialogue			Practice		Rating
	AI reply time (seconds)	Student messages	AI tutor messages	Concepts shown	Practice shown	Practice credit (%)	Helpfulness (1–5)
Gemini 3.1 Pro (high)	6.77	34.7 7/11	67.8	3.72	23.4	94.9%	4.12
Gemini 3.5 Flash (low)	3.61	32.9	65.8	3.04	22.8	94.8%	4.33 6/11
Gemini 3.6 Flash (low)	2.28	30.4	70.4	3.84	27.0	90.3%	4.05
GPT-5.4 mini (off)	1.88 11/11	29.8	85.6 11/11	4.45	37.0 11/11	85.9%	3.73
Gemini 3.7 Flash (med)	2.46	28.1	63.1	3.69	24.5	93.4%	4.23
Gemma 4 31B (high)	13.93	24.5	41.2	3.69	20.5	91.6%	3.74
Opus 4.8 (x-high)	11.05	20.9	52.7	4.96	23.8	92.9%	4.11
Opus 4.8 (off)	11.03	20.0	53.2	4.95	24.7	93.1%	4.15
GPT-5.5 (high)	7.61	18.9	56.4	4.71	28.1	95.1% 5/11	3.99
Opus 5 (high)	14.63	17.2	48.0	5.38 11/11	23.2	91.9%	4.01
Kimi K2.6 (med)	19.43	17.1	40.2	4.36	14.3	88.0%	3.88
GPT-5.5 Pro (med)	30.99	13.5	38.6	4.70	18.8	94.3%	3.76

Color: pale=last rank; dark=first rank within each column. Outlined cells: leader; resolved wins / rivals.

Figure D.4: **The tutoring experience across AI tutors.** Combined estimates of pace, dialogue, practice and student ratings for the 12 shared AI tutors. Fractions count how many of the other 11 tutors have significantly longer reply times or lower values on the other measures than the column leader ($p<.05$). The practice outcome is the percentage of distinct answered problems receiving final platform credit. See Section D.2 for the process comparisons and Appendix D.1 for definitions and adjustment.

The cost-per-gain measure is $\bar{c}_i/\bar{g}_i$, the mean session cost divided by mean learning gain for AI tutor i, as in Section 6 and Figure 5, summarizing the cost required for each unit of learning gain for each tutor. A paired bootstrap resamples session records within each AI tutor, keeping gain and cost together (10,000 draws). A bootstrap draw is valid if the mean cost is finite, meaning a cost is available, and the mean gain is positive. Intervals span the 2.5th and 97.5th percentiles when at least 95% of draws are valid, and the plots use unrounded values. The Reproducibility section describes code availability.

AI reply time in Figure 4B, C is the median recorded provider-call duration within each session, then the median across sessions for each AI tutor. This measures the duration of the model call, rather than time to the first streamed text or the full delay experienced in the browser. Engagement is averaged across sessions for each AI tutor. In Figure 4A, B, an AI tutor is on the frontier if no other AI tutor offers equal or higher learning gain at equal or lower cost or reply time, with at least one strict improvement. This identifies the best cost-time-gain trade-off in the sample. Frontier membership is based on sample estimates and may change with sampling variation.

After tutoring and before the post-test, AI students answered "How helpful was your tutor conversation for your learning?" on a five-point scale. Its endpoints were 1, "Not at all helpful," and 5, "Extremely helpful." All 2,139 AI sessions have a response; the figures show mean ratings within each AI tutor, summarizing perceived helpfulness of the tutor conversations.

In Figure D.4, the practice credit percentage is 100 times the number of distinct problems receiving final credit divided by the number of distinct answered problems. Final credit uses the AI tutor's answer key and includes automatic credit for non-blank open responses. This gives the share of answered problems that were ultimately credited as practice.

Reply time and student-message correlations use one observation per AI tutor: 13 per section and 12 in Combined. Spearman correlations compare the ranks of median full-reply time and mean student-message count. The six-test Holm correction covers Pearson and Spearman correlations in Quantitative, Verbal, and Combined; only Spearman results are reported. Figure 4C shows the combined Spearman result on a linear reply time axis, with an unweighted least-squares fit of messages against $\log_{10}$ reply time. Its pointwise 95% confidence band for the fitted mean uses residual variance and a Student-t distribution with 10 degrees of freedom. Figure D.3 uses the same

log-time fit and displays Spearman correlations and their p values for Quantitative, Verbal, and Combined, supporting Section 5.

The interaction map in Figure D.4 compares seven session measures after adjustment for centered pre-test score and, in Combined, the GRE section. Pairwise comparisons use HC3 covariance. We log-transform latency and use $\log(1 + x)$ for student messages, tutor messages, and problems shown. Concept counts, the percentage of problems receiving final credit, and helpfulness ratings remain on their original scales. Shorter reply times and higher values on the other measures rank first.

Table D.1 reports associations of cost, reply time, and student messages with mean learning gain across AI tutors. The dialogue and practice analyses in Section D.2 instead compare individual sessions. Each AI tutor contributes one observation: 13 per section and 12 shared AI tutors in Combined. Combined retains each AI tutor's proportions of Quantitative and Verbal sessions. Spearman coefficients describe correlations between ranks.

Table D.1: Associations of cost, reply time and student engagement with mean learning gain across AI tutors, corresponding to Figures 4A, B and D.3. See Appendix D.1 for cost, reply time and engagement definitions. k is the number of AI tutors.

Predictor	Scope	k	Spearman ρ
Cost	Quantitative	13	0.462
Cost	Verbal	13	0.335
Cost	Combined	12	0.378
Reply time	Quantitative	13	0.198
Reply time	Verbal	13	−0.159
Reply time	Combined	12	−0.007
Student messages	Quantitative	13	0.110
Student messages	Verbal	13	0.203
Student messages	Combined	12	0.203

D.2 Dialogue, practice, and learning

Beyond measuring equivalence and performance, we seek to understand which AI tutor characteristics are associated with learning gains. In an exploratory analysis, learning gains were associated with doing more correct practice, which was associated with greater student engagement. These findings linked learning gains to student activity, but failed to connect learning gain to a characteristic of the AI tutor.

Here, the strong association we found earlier between AI reply time and student engagement (Figure 4C) supplied that connection. We examined this chain in 1,137 Quantitative AI-tutoring sessions using regressions accounting for pre-test score, AI tutor, and assessment form. Faster replies were associated with more student engagement, more engagement with more correct practice, and more correct practice with greater learning gains (Figure D.5; all $p \leq .002$). These associations were not statistically significant in Verbal. Verbal students sent fewer, longer messages with similar word totals, suggesting message count may capture engagement differently across sections. The analysis details follow below and in Table D.2.

A distinctive feature of StudentBench is that model latency is a first-class citizen. In most AI benchmarks, latency is measured as a secondary outcome. In StudentBench, latency, through engagement and practice, may indirectly influence the primary outcome of learning gain. We suspect low latency and its associated engagement and practice may help explain why models that perform poorly on the teaching-capability leaderboards in Figure 3 still achieve human-level learning gains on the GRE.

Figure D.5 uses 2,135 AI-tutoring sessions: 1,137 Quantitative and 998 Verbal. Four of the 2,139 sessions lack a recorded first answer for at least one practice problem; the same complete set is used for all three associations. Reply time is the mean positive recorded duration of complete AI provider calls within a session, including blank-content replies. Figures 4B, C and D.3 instead use medians to summarize reply time by AI tutor. Engagement and correct practice are defined in Section 5.2.

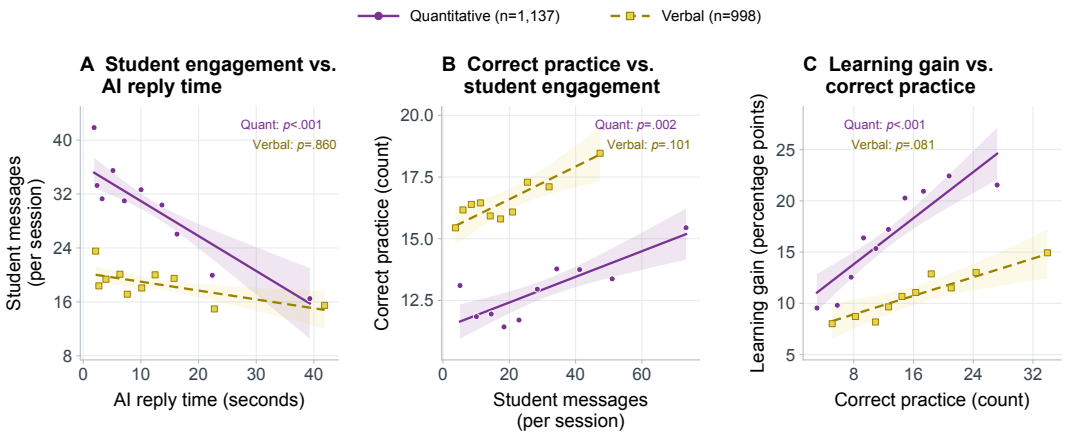

Figure D.5: **Explaining learning gains.** Dots show averages for ten similarly sized groups of sessions per GRE section, ordered by AI reply time (**A**), student messages (**B**), or correct practice (**C**). Values on the vertical axis and regression lines adjust for pre-test score, AI tutor, and assessment form. Lines use individual sessions (1,137 Quantitative and 998 Verbal); shading shows pointwise 95% confidence intervals for the fitted means. Tests account for 276 exploratory comparisons (Table B.1). See Section 5.2 for definitions.

Practice credit is recorded against AI-authored answer keys; students may consult the tutor before submitting an answer.

The metric comparison was exploratory. We fixed two reply time measures (mean and median), three engagement measures (student messages, messages with at least five whitespace-delimited tokens, and total student tokens), and five practice measures (distinct answered problems, submissions including retries, final all-format credit, final closed-response credit, and correct practice). Their associations with one another and learning gain gave 78 tests across Quantitative, Verbal and Combined on the common 2,135 sessions. Another 78 tests used $\log(1 + x)$ for latency and counts, and 42 used messages of at least ten tokens or nonwhitespace character counts. Learning gain and pre-test score were not transformed. The displayed definition was selected after this comparison; complete results for every definition are supplied with the data.

A fixed 30-minute split added 78 tests on 2,133 sessions: early reply time with late engagement, early engagement with late practice, and late practice with learning gain. Windows were $[0, 30)$ and $[30, 60)$ minutes from the unique server intervention start. Early calls had an estimated start and recorded completion within the first window; start was estimated by subtracting duration from reply-persistence time. Practice timestamps record events after feedback returns. All p values for these analyses use Holm correction across the resulting 276 tests (Table B.1).

Regressions use ordinary least squares with HC3 covariance and a normal reference distribution, adjusting for linear pre-test score, AI tutor version and assessment form; Combined additionally distinguishes section-specific forms. Predictors are centered and divided by their sample standard deviation within each analysis subset. Table D.2 rescales coefficients and intervals to effects per ten seconds, student messages or correct practice problems. Gains are percentage-point changes on unaided assessments.

Fitted lines in Figure D.5 average predictions across sessions within each section, holding each session's pre-test score, AI tutor and assessment form at their recorded values. Pointwise 95% confidence bands for the fitted mean use the average design vector and HC3 covariance, with normal critical value 1.96. Plotted points show means for ten similarly sized groups ordered by the predictor. Outcomes subtract the fitted contribution of the other covariates and add back their section-average contribution. Lines and bands run between the lowest and highest group means. Every included session contributes to a group and to its section's fit.

The Quantitative and Combined associations retained their directions under log transformations, partial-rank checks, fixed 1st/99th-percentile winsorization, all-available-row fits and omission of each AI tutor version. Excluding repeat participants left 1,985 sessions and preserved all seven

selected Combined associations. The temporal comparisons were less consistent: early engagement with late correct practice had p=.091 in Quantitative, and late correct practice with gain had p=1.00 in Verbal. Complete estimates and diagnostics are included in the accompanying data.

Table D.2: **Student engagement, correct practice and learning.** Selected exploratory associations in Figure D.5, using 1,137 Quantitative, 998 Verbal and 2,135 Combined sessions. See Section 5.2 for the definition of correct practice. Coefficients are outcome changes per ten seconds, student messages or credited problems; gain is in percentage points. Intervals are pointwise 95% HC3 intervals; p values follow Table B.1.

Section	Predictor / outcome	β_{10}	95% CI	p
Quant.	Reply time / student messages	-5.19	$[-7.09, -3.29]$	$< .001$
Quant.	Student messages / correct practice	0.52	$[0.29, 0.75]$	.002
Quant.	Correct practice / learning gain	5.62	$[3.97, 7.26]$	$< .001$
Verbal	Reply time / student messages	-1.31	$[-2.24, -0.39]$	.860
Verbal	Student messages / correct practice	0.66	$[0.29, 1.04]$	.101
Verbal	Correct practice / learning gain	2.27	$[1.00, 3.54]$	.081
Combined	Reply time / student messages	-3.46	$[-4.42, -2.51]$	$< .001$
Combined	Student messages / correct practice	0.53	$[0.33, 0.72]$	$< .001$
Combined	Correct practice / learning gain	3.82	$[2.82, 4.82]$	$< .001$

E Expert review of lesson plans and practice problems

E.1 Common rubric and grouping

Following the learning gain results, expert reviewers graded pairs of lesson plans using a rubric with 8 criteria. The rubric, seen in Table E.1, was used for both Quantitative and Verbal, and the criteria can be split into two groups: lesson planning and practice-problem creation and design. The division into the two groups is an exploratory interpretation of the rubric. Note that for Verbal, the appropriate difficulty asks if the concepts are appropriate for the student.

Table E.1: Eight shared criteria for expert review of teaching materials. Letters A–H correspond to Figure E.1. Figure 3A pools preference ratings for A–D and H (lesson planning); Figure 3B pools E–G (practice-problem creation and design). See Section 4.1 for evaluation details.

Criterion	What the reviewer evaluates
A. Relevant concepts	Choosing concepts that address the student's pre-test errors.
B. Concept grouping	Teaching related topics together.
C. Concept prioritization	Placing high-impact concepts earlier.
D. Time allocation	Dividing time according to errors, difficulty and likely benefit.
E. Practice alignment	Matching practice to the student's errors and skill level.
F. Appropriate difficulty	Choosing problems or concepts suitable for what the student knows.
G. Example accuracy	Providing correct problems, solutions and answer keys.
H. Test-taking strategies	Helping the student handle GRE formats and time-sensitive decisions.

The comparison uses a student's pre-test and the lesson plan an AI tutor created for the session, and then has another AI tutor generate a plan based on the pre-test, thus copying the production environment exactly. The AI tutors for the comparison plans were selected to meet quotas for each pair of AI tutors based on the available plans. The alternative, or comparison, plans were generated for expert review and did not replace the original plans during tutoring. Of the 381 student pre-tests, 340 came from the learning cohort and 41 from supplementary or excluded cohorts. Plans for six Quantitative pre-tests used a richer prompt, contributing 16 reviews with the same prompt within each pair. Expert preferences therefore cover a broader set of student pre-tests than the learning leaderboard.

Table E.2: **Expert-ranking sensitivity to repeated reviewers.** This checks the expert comparisons in Section 4.1. Counts show how many version-specific AI tutor entries differ significantly from Opus 5 ($p<.05$). The original analysis clusters uncertainty by pre-test; the sensitivity clusters by reviewer and pre-test. Fitted scores are held fixed. Ratings marked as ties are excluded.

Evaluation	Ratings	Pre-tests	Original	Sensitivity
Lesson planning	9,577	381	9/12	8/12
Practice creation/design	5,265	380	9/12	8/12
All eight shared criteria	14,842	381	9/12	8/12

The expert reviewers were recruited for GRE experience at ETS or Kaplan, or at least five years of tutoring experience. The expert reviewers and live human tutors were separate groups, with no overlap. Reviewers saw the student's graded pre-test and two plans labeled A and B, with AI tutor names hidden and A/B order randomized for each reviewer. They rated each plan from 1 to 5, gave a five-level relative preference on the 8 criteria, and explained their decision. A queue balanced coverage of pairs of AI tutors; the comparisons have one to five distinct reviewers each.

The 2,028 reviews comprise 1,136 Quantitative and 892 Verbal reviews, covering 852 distinct comparison tasks. Of the 51 reviewers, 39 reviewed Quantitative plans and 38 reviewed Verbal plans; some reviewed both. The common eight criteria yield 16,224 criterion-level evaluations: 10,140 for lesson planning and 6,084 for practice-problem creation and design. The lesson-plan comparisons cover 12 AI tutors with 13 version-specific entries; Gemini 3.6 Flash has no expert-review estimate.

For AI tutors a and b, the Bradley–Terry model (Turner & Firth, 2012) estimates preference probability $\text{logit}^{-1}(\theta_a - \theta_b)$. Ties are excluded; slight and strong preferences receive weights one and two, respectively. A pre-test is counted only if it contributes at least one preference not marked as a tie to the ranking. All 381 student pre-tests contribute to the lesson-planning ranking, but only 380 contribute to the practice ranking because reviews on one pre-test yielded only ties on the practice criteria. The plotted intervals account for repeated evaluations based on the same student pre-test. Appendix E.2 also allows for experts reviewing multiple pre-tests. Ranking scores reflect fitted expert preference.

The Quantitative test-taking-strategy criterion asked about GRE-format weaknesses in 590 reviews and broader test-taking strategy in 546. Both wordings map to the shared strategy criterion; the supplementary rubric preserves each version. The AI tutors were not all evaluated on the same student pre-tests or against the same alternatives; the ranking is based on the available comparisons.

E.2 Repeated expert reviewers

The 51 experts contributed 2,028 reviews; 39 reviewed more than one student pre-test. A sensitivity check of pooled planning, practice and eight-criterion comparisons uses uncertainty clustered by both reviewer and pre-test (Cameron et al., 2011). The sensitivity analysis's results are in Table E.2, and the analysis leaves the fitted scores, preference weights, and tie exclusions unchanged. It adds the reviewer- and pre-test-clustered covariance matrices and subtracts their intersection, using the usual finite-cluster correction for each component. Tests use a Student-t reference with 50 degrees of freedom.

After accounting for repeated reviewers, Opus 5 no longer differs significantly from Opus 4.8 (off) in planning or the eight-criterion ranking, or from GPT-5.5 in practice design.

E.3 Overall and criterion-specific leaderboards

Figure E.1 expands the two leaderboards in Figure 3A, B into the eight criteria in Table E.1 and overall rankings. Panels A–H fit each criterion separately; panels I–K fit the pairwise ratings from all eight criteria together for Quantitative, Verbal and both sections combined. Each section includes 12 AI tutors and 66 pairs; the combined fit includes 13 version-specific AI tutor entries and 78 pairs. Scores are centered within each fit, so values from different panels are not directly comparable.

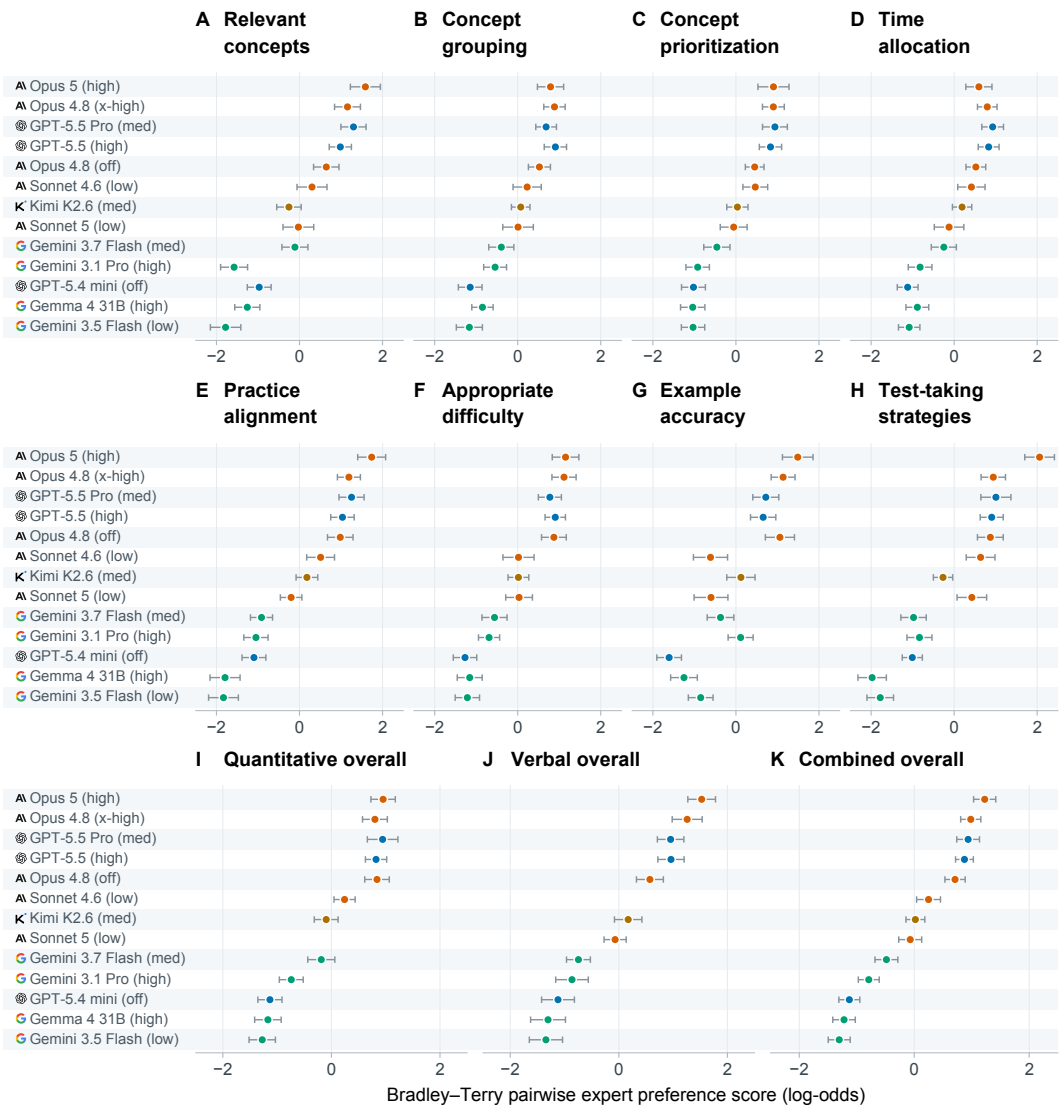

Figure E.1: **Expert preference by teaching criterion and overall.** A–H, the eight criteria in Table E.1; I–K, all eight fitted jointly for Quantitative, Verbal and both sections combined, respectively. These rankings expand Figure 3A, B. Points show centered Bradley–Terry estimates; whiskers are pointwise 95% intervals clustered by student pre-test. For each panel, we report the number of student pre-tests and expert ratings. Ratings marked as ties are excluded. A, 371 pre-tests and 1,899 ratings; B, 370 pre-tests and 1,905 ratings; C, 372 pre-tests and 1,916 ratings; D, 376 pre-tests and 1,956 ratings; E, 377 pre-tests and 1,939 ratings; F, 365 pre-tests and 1,653 ratings; G, 366 pre-tests and 1,673 ratings; and H, 378 pre-tests and 1,901 ratings. Overall fits use 246, 135 and 381 pre-tests, with 8,206, 6,636 and 14,842 ratings, respectively. All panels share the combined AI tutor order; blank cells mean not evaluated. Scores are comparable only within panels.

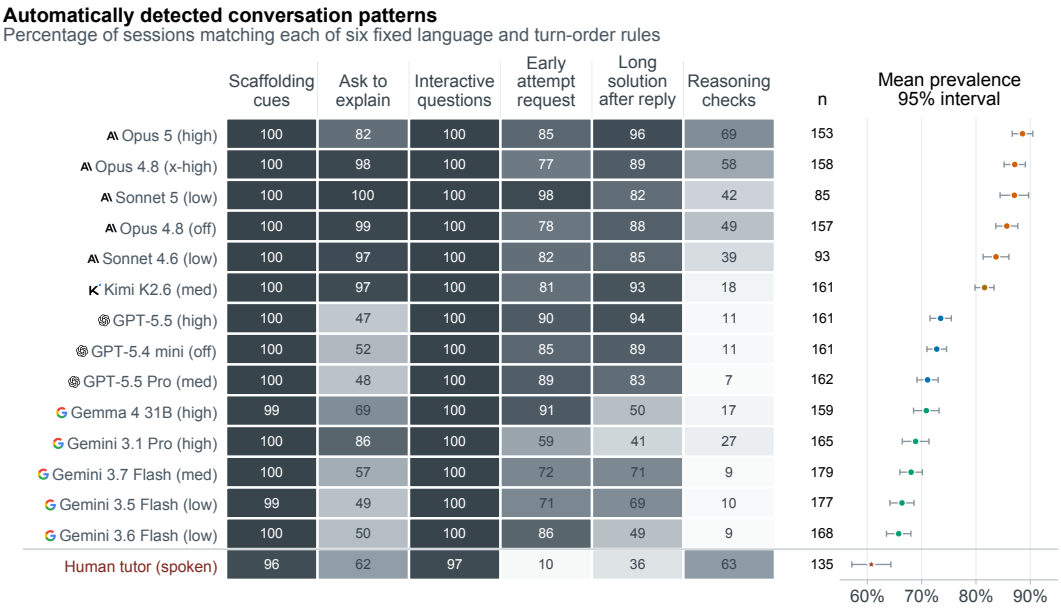

Figure F.1: **How tutors interact with students.** Cells show the percentage of sessions matching each rule underlying the ranking in Figure 3C. Dots show mean prevalence across the six indicators, with 95% Student-t intervals across sessions; n is the transcript count. The human reference uses spoken sessions, whereas AI sessions use typed chat. See Figure F.2 for Quantitative and Verbal results and Table F.1 for indicator definitions.

F Conversational pedagogy and student-reported problem quality

Transcripts cover all 2,139 AI sessions and 135 of the 140 human sessions. Human transcripts group consecutive lines from the same speaker into turns and include only the period from the start of tutoring to the transition to the post-test. Recordings failed for five human sessions. The conversational-pedagogy evaluation uses six rules of "good" tutoring synthesized from established literature in the learning and cognitive sciences (Wood et al., 1976; VanLehn, 2011; Chi et al., 1994; Chi & Wylie, 2014; Kapur, 2008; Koedinger & Aleven, 2007; Sweller & Cooper, 1985; Dunlosky et al., 2013; Renkl, 2014) (see Table F.1). We then identified the prevalence of these rules using deterministic algorithms that extract wording and turn order to identify scaffolding cues, requests for explanation, interactive questions, early requests to attempt a problem, long solutions after a student reply, and reasoning checks. For each tutor, each rule's prevalence is the percentage of transcripts in which it appears. We report the mean prevalence across the six rules.

Although these rules are based on prior literature, we did not validate them as predictors of learning lift in this study. Additionally, human sessions used spoken conversation over video calls, whereas AI sessions used typed chat. The same rules may behave differently in spoken and typed conversations; future evaluations may use a model's voice mode when available.

Table F.1: Operational definitions for the six transcript indicators in Figures 3C, F.1 and F.2.

Indicator	Detection rule
Scaffolding cues	A tutor response following a student message of at least two words contains evaluative language, or shares at least two substantive words with that message and contains a scaffolding cue.
Request for explanation	At least one tutor question contains a request for reasoning or explanation.

Indicator	Detection rule
Interactive questions	At least four tutor messages contain a question mark or an invitation to respond.
Early attempt request	At least one of the first five tutor messages asks the student to try or attempt a problem.
Long solution after a reply	After a student message of at least two words, a tutor message of at least 140 words includes a solution or calculation cue.
Reasoning checks	At least two tutor questions contain a reasoning-check or strategy cue.

A Quantitative

Sessions matching each conversation rule (%)

	Scaffolding cues	Ask to explain	Interactive questions	Early attempt	Long solution after a reply	Reasoning checks	n
Opus 5 (high)	100	63	100	91	99	93	75
Opus 4.8 (x-high)	100	97	100	81	95	60	86
Opus 4.8 (off)	100	98	100	80	95	56	84
Kimi K2.6 (med)	100	95	100	88	98	24	85
Sonnet 4.6 (low)	100	97	100	82	85	39	93
Gemma 4 31B (high)	99	63	100	95	48	11	87
GPT-5.4 mini (off)	100	19	100	89	96	8	90
Gemini 3.7 Flash (med)	100	34	100	97	72	2	87
Gemini 3.6 Flash (low)	100	35	100	89	66	9	89
GPT-5.5 (high)	100	6	100	94	98	0	89
Gemini 3.5 Flash (low)	100	25	100	76	80	3	97
Gemini 3.1 Pro (high)	100	77	100	65	37	2	91
GPT-5.5 Pro (med)	100	5	100	88	87	0	86
Human tutor (spoken)	95	57	98	13	45	60	60



B Verbal

Sessions matching each conversation rule (%)

	Scaffolding cues	Ask to explain	Interactive questions	Early attempt	Long solution after a reply	Reasoning checks	n
Sonnet 5 (low)	100	100	100	98	82	42	85
Opus 5 (high)	100	100	100	79	94	45	78
Opus 4.8 (x-high)	100	100	100	72	82	56	72
Opus 4.8 (off)	100	100	100	77	79	41	73
GPT-5.5 (high)	100	97	100	85	89	24	72
GPT-5.5 Pro (med)	100	97	100	89	78	14	76
Kimi K2.6 (med)	100	99	100	74	88	12	76
GPT-5.4 mini (off)	100	94	100	80	80	14	71
Gemini 3.1 Pro (high)	100	97	100	53	46	57	74
Gemma 4 31B (high)	100	75	100	85	51	24	72
Gemini 3.5 Flash (low)	99	79	100	65	55	18	80
Gemini 3.7 Flash (med)	100	78	100	48	70	15	92
Gemini 3.6 Flash (low)	100	67	100	84	30	9	79
Human tutor (spoken)	96	67	96	8	29	65	75

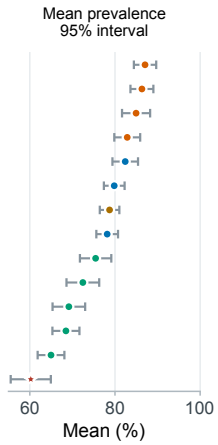

Figure F.2: **Conversational pedagogy by GRE section. A,** Quantitative. **B,** Verbal. Panels expand the Combined ranking in Figure 3C using the six rules in Table F.1. Cells give the percentage of sessions satisfying each rule; n counts available transcripts. The right column shows the mean of the six indicators with 95% Student-t intervals across sessions. Human tutoring is a spoken reference and AI tutoring uses typed chat.

F.1 Student-reported answer disputes

Students could press "I disagree with this answer — continue" after receiving feedback on a practice answer. For each AI tutor, the dispute rate is the fraction of distinct answered problems with at least

one recorded dispute. Repeated clicks on one problem count once. The button records disagreement with the marked answer, not an independently verified error in the problem or answer key.

Figure F.3 records 523 disputed answers among 21,769 answered Quantitative problems and 120 among 23,694 Verbal problems. All observed Verbal rates were below one dispute per 100 answered problems. The 95% percentile intervals use 4,000 session resamples, with ratios recomputed from each resample's totals. A zero observed count gives a zero-width interval but does not establish zero risk. Kruskal–Wallis tests compare per-session rates among AI tutors; the Quantitative, Verbal and Combined tests were significant ($p<.001$).

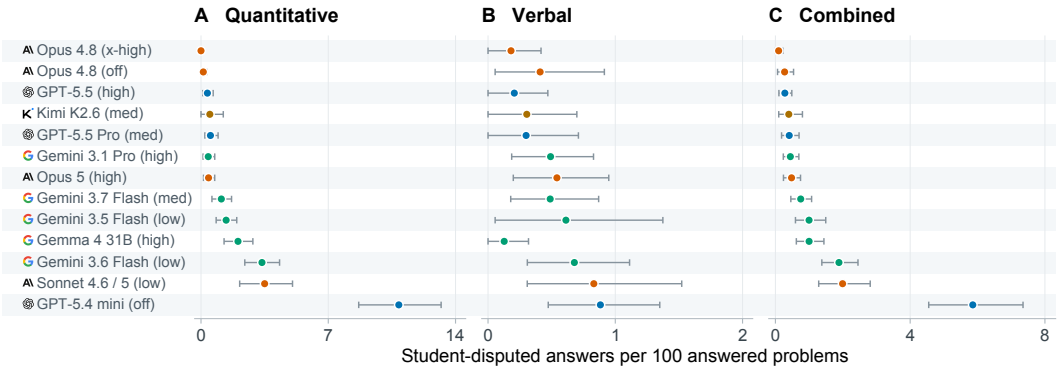

Figure F.3: **How often students disputed a practice answer.** A–C show flagged problems per 100 distinct answered problems in Quantitative, Verbal and Combined. Whiskers are 95% session-bootstrap intervals across 1,139 Quantitative, 1,000 Verbal and 2,139 Combined sessions. Horizontal scales differ to show the smaller Verbal rates. Rows share the Combined ordering. The Sonnet row pools its section-specific versions in Combined. These student reports complement the expert practice rankings in Figure 3B; they record student disagreements with answers, not independently verified errors.

F.2 Agreement with expert judgments

Higher expert preference for practice-problem creation and design (Figure 3B) was associated with fewer student disputes in Quantitative and Combined. Correlations compare AI tutor ranks for practice-problem creation and design (criteria E–G) and, separately, example accuracy (G) with the negative of each AI tutor's dispute rate. Positive correlations therefore indicate agreement between expert preferences and fewer student disputes. Each section includes 12 AI tutors with both expert ratings and student-dispute data; Combined includes 11, excluding the row that pools the two Sonnet versions. Gemini 3.6 Flash has no expert-review estimate. The section-specific expert scores are refitted from their original pairwise ratings using the same criteria as the Combined ranking.

For practice-problem creation and design, Spearman correlations were ρ=0.713 in Quantitative (p=.035), ρ=0.315 in Verbal (p=.388), and ρ=0.773 in Combined (p=.033). For example accuracy, the corresponding correlations were 0.895 ($p<.001$), 0.406 (p=.388), and 0.782 (p=.033). These exploratory two-sided tests use 99,999 random permutations of AI tutor labels. They treat the estimated AI tutor ranks as fixed and do not account for uncertainty in those ranks. Expert review concerns examples in plans, whereas student disputes can also concern practice generated later in tutoring.

G Prompt-design comparisons

The learning results in the main text and in particular in Figure 1 use the fixed main-study prompts shown in Appendix A.5. However, we did an early Quantitative pilot containing 326 sessions from 312 participants across Opus 4.8, Gemini 3.5 Flash and GPT-5.5 focusing on the prompt design.

MINIMAL AND EXPANDED PROMPTS PILOT

The pilot compared five combinations of minimal, intermediate and expanded lesson-plan and tutor prompts (Table G.1). When both prompts were minimal, the pooled mean gain was 14.27 percentage points, compared with 8.85 when both were expanded. Mean gains were higher with minimal prompts for each AI tutor (Figure G.1A). These pilot data informed the main-study prompts and the data points are excluded from the main learning cohort.

The numbers of sessions differed across prompt conditions and AI tutors, and the lesson-plan generator differed from the chat tutor in 41 sessions. An exploratory regression adjusted for pre-test score, assessment form and the AI tutor used for planning, with a full interaction between AI tutor and prompt condition and covariance clustered by first-attempt participant identity. The joint interaction was inconclusive (χ^2_8=10.00, p=.265). None of the three tutor-specific expanded/expanded minus minimal/minimal contrasts, or their three pairwise differences, was significant.

Separate cohorts studied at different times used expanded prompts with Opus 5 and Gemini 3.6 Flash; the main-study cohorts used minimal prompts. Mean learning gains were higher with the expanded prompts for both AI tutors: 18.79 versus 15.31 percentage points for Opus 5 and 19.22 versus 14.65 for Gemini 3.6 Flash (Figure G.1B). After adjustment for baseline and form, the corresponding differences were 3.61 percentage points (95% CI $[-1.66, 8.89]$; p=.359) and 4.62 percentage points ($[-0.17, 9.42]$; p=.177). Their interaction was also inconclusive (p=.781). The comparisons use HC3 covariance. Because the cohorts were studied at different times, these comparisons do not isolate the effect of the prompts.

Table G.1: Prompt-selection evidence for the fixed-instruction design in Section 2. All five Quantitative pilot settings are shown; these sessions are separate from the main learning results in Figure 1. Cells give mean percentage-point gain, with the number of sessions in parentheses. The two terms in each row identify the lesson-plan and tutor prompts, respectively. AI tutor columns identify the deployed chat tutor; the lesson-plan generator differed in 41 sessions.

Lesson plan / tutor	Opus 4.8	Gemini 3.5 Flash	GPT-5.5	All
Minimal / Minimal	15.23 (9)	17.78 (10)	13.56 (62)	14.27 (81)
Minimal / Expanded	14.81 (27)	8.50 (27)	12.65 (36)	12.06 (90)
Intermediate / Intermediate	11.62 (29)	2.06 (9)	11.40 (26)	10.19 (64)
Expanded / Minimal	8.38 (19)	2.65 (7)	23.23 (11)	11.71 (37)
Expanded / Expanded	8.02 (18)	11.85 (10)	8.26 (26)	8.85 (54)

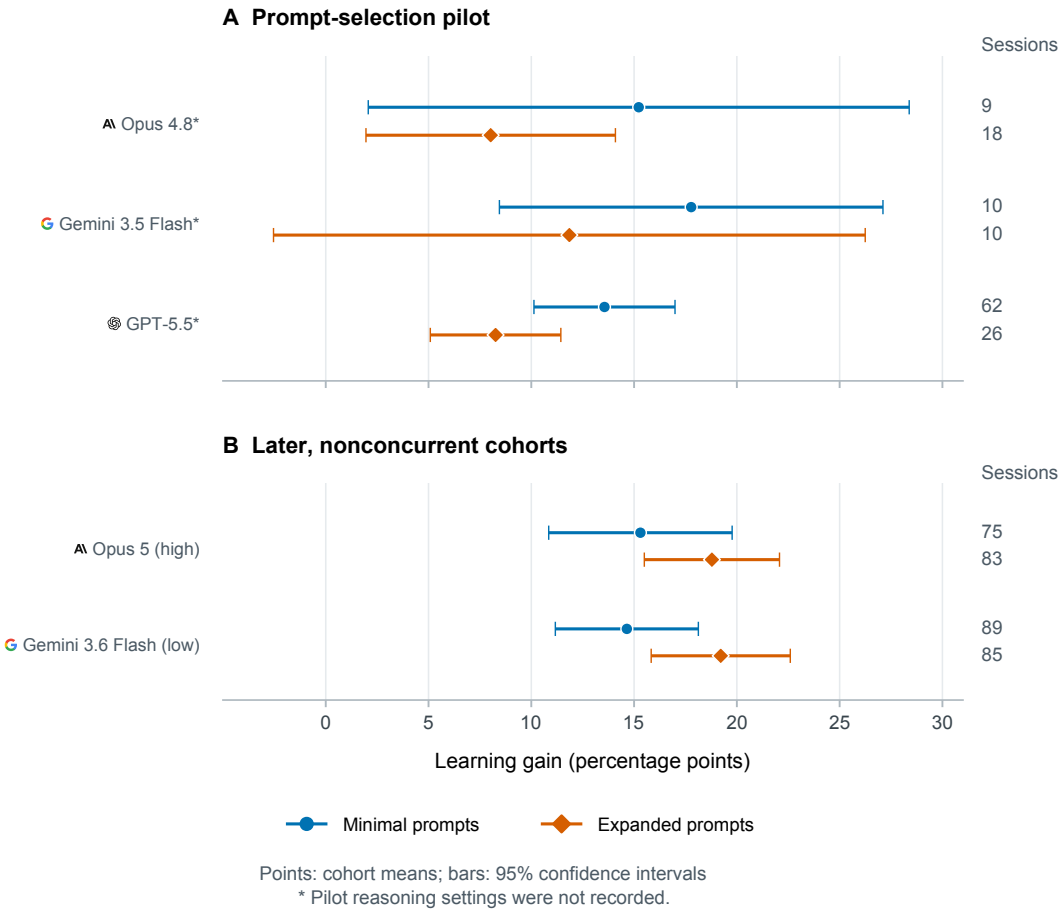

Figure G.1: **Learning with minimal and expanded prompts.** Supporting the prompt design in Section 2, points show mean Quantitative learning gain with separate Student-t 95% confidence intervals; counts on the right show the number of sessions. Both lesson-plan and tutor prompts are minimal for blue circles and expanded for orange diamonds. **A,** Three AI tutors in the exploratory prompt-selection pilot. This panel shows the two matched prompt settings; see Table G.1 for all five settings, including mixed and intermediate prompts. **B,** Two later AI tutors, comparing supplementary cohorts that used expanded prompts with main-study cohorts that used minimal prompts. Differences are descriptive: the pilot informed prompt selection, and the later prompt conditions were not concurrently randomized.