LearnLM: Improving Gemini for Learning

LearnLM Team, Google

Today's generative AI systems are tuned to present information by default rather than engage users in service of learning as a human tutor would. To address the wide range of potential education use cases for these systems, we reframe the challenge of injecting pedagogical behavior as one of *pedagogical instruction following*, where training and evaluation examples include system-level instructions describing the specific pedagogy attributes present or desired in subsequent model turns. This framing avoids committing our models to any particular definition of pedagogy, and instead allows teachers or developers to specify desired model behavior. It also clears a path to improving Gemini models for learning—by enabling the addition of our pedagogical data to post-training mixtures—alongside their rapidly expanding set of capabilities. Both represent important changes from our initial tech report [1]. We show how training with pedagogical instruction following produces a LearnLM model (available on Google AI Studio) that is preferred substantially by expert raters across a diverse set of learning scenarios, with average preference strengths of 31% over GPT-4o, 11% over Claude 3.5, and 13% over the Gemini 1.5 Pro model LearnLM was based on.

1. Introduction

Our initial tech report [1] from May 2024 surveyed the history and current landscape of education technology, discussed the potential impact of generative artificial intelligence (gen AI) on education, and presented our collaborative approach to developing evaluations.

Following its publication, we received input from across the international education sector, including schools, educational technology ("EdTech") companies, non-profit organizations, and government agencies eager to try our models or otherwise collaborate. Through review of these submissions, over 20 follow-up interviews, and input from Google product teams building gen AI powered learning features, we can summarize the key findings as follows:

1. Pedagogy[1], or rather, ideal behavior of an AI tutor, is prohibitively difficult to define given the wide range of grade-levels, subjects, languages, cultures, product designs, and philosophies that must be accommodated. While there are many commonalities, appropriate behavior in different contexts may be different or even contradictory, and it is best left to the developer or teacher to specify.
2. In developing AI learning systems, the most commonly cited, immediately useful behavior in an underlying model is the ability to follow system instructions to create interactive tutor-led exercises. Teachers or developers who specify these instructions want to feel confident that the AI tutor will follow the specified instructions accurately, even if a student tries to circumvent them (e.g., "do not give away the answer" or "stay on topic").
3. Post-hoc fine-tuning for each application can be effective in the short-term, but is impractical because of cost, maintenance, and rapidly improving base models. Thus, despite its shortcomings, prompting will likely remain the best way for education product developers to specify behavior.

This paper describes how we have updated our modeling and evaluation methodology in light

[1]We use the term *pedagogy* in as broad a sense as possible, certainly not limited to children, to evoke techniques of teaching and associated learning by humans.

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

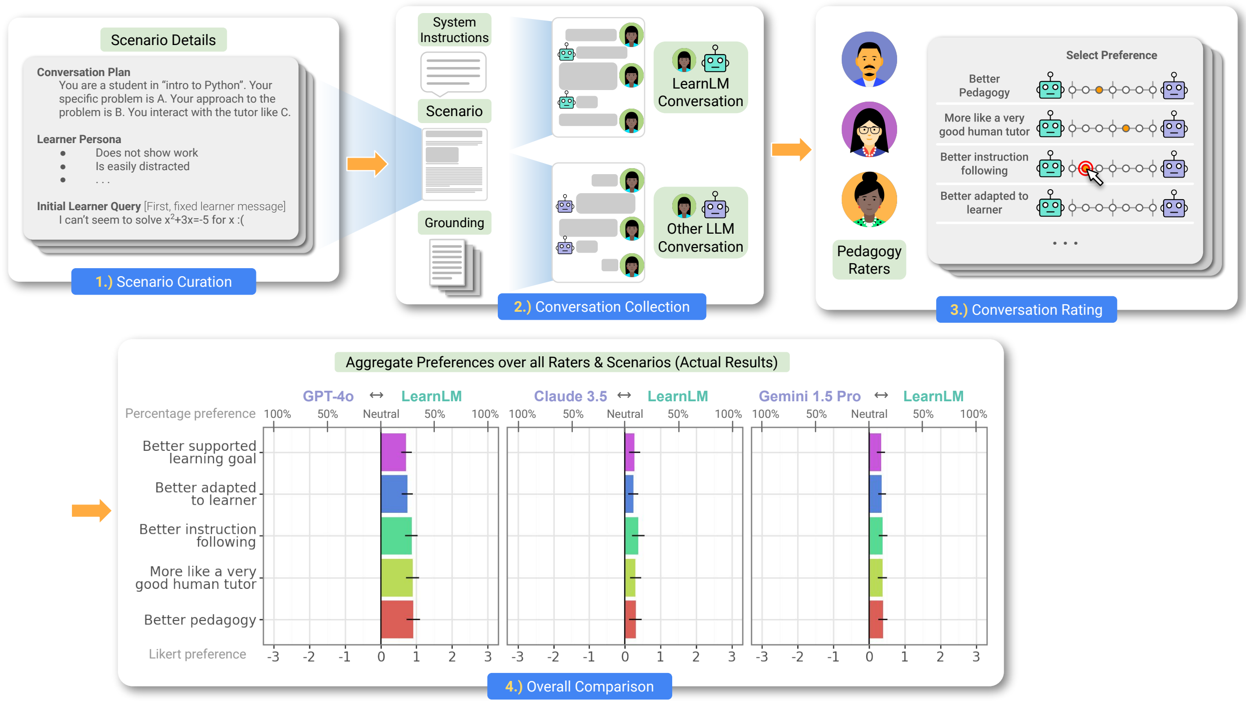

Figure 1 | An overview of our three-stage human evaluation pipeline and our results for comparing LearnLM with other systems. (1) Learning scenarios are developed that allow raters to role-play specific learners interacting with pairs of AI tutors (2). Grounding material (e.g. an essay, homework problem, diagram, etc.) and System Instructions specific to each scenario are passed as context to each model. The resulting conversation pairs are reviewed by pedagogy experts (3) who answer a range of questions assessing each model on its own as well as their comparative performance. These comparative ratings (on a seven-point -3 to +3 Likert scale) are aggregated (4) to show overall preference for LearnLM over GPT-4o, Claude 3.5, and Gemini 1.5 Pro. See Section 4 for more detailed results.

of these observations. Specifically, we cast our work as *pedagogical instruction following*, meaning that we contextualize training and evaluation examples with system-level instructions that describe desired pedagogical behavior. This approach avoids any narrow specification of how systems should behave, and allows us to effectively add pedagogical data to the rest of Gemini's training mixture without conflicts of persona or style. We also include Reinforcement Learning from Human Feedback (RLHF) [2] in our training procedure to allow models to follow more nuanced pedagogical instructions and preferences.

Using the updated methodology, we trained a new version of *LearnLM*, which is based on Gemini 1.5 Pro[2] [3]. In our evaluations against contemporaneous flagship models, each representing a company's premier offering as of 2024-10-01, expert pedagogical raters preferred LearnLM with an average preference strength of 31% over GPT-4o, 11% over Claude 3.5 Sonnet, and 13% over the original Gemini 1.5 Pro (see Figure 1). LearnLM is available as an experimental model on Google AI Studio along with documentation of example use cases and suggested prompts. We welcome any feedback, which will help inform our future research and prioritization. As we improve LearnLM for teaching and learning, we are also working to bring these advances into Gemini models, so any developers using Gemini can benefit from the improvements made via LearnLM research.

Section 2 describes how we trained LearnLM for pedagogical instruction following and Section 3 explains how we updated our scenario-based evaluation design accordingly. Section 4 shows a detailed analysis of results comparing LearnLM with other premier model offerings. Finally, Section 5 outlines some future work, especially with regards to continued evaluation. Training and evaluation

[2]Specifically gemini-1.5-pro-002 (release notes).

of LearnLM is done across a broad range of core academic subjects, and we include a feasibility study on medical education subjects in Appendix C.

2. Modeling

In our original tech report [1], we adapted the behavior of a base model by Supervised Fine-Tuning (SFT) with a range of synthetic and human-written datasets. Since then, we have made a number of substantial changes to our training strategy: First, we updated our SFT data according to our focus on pedagogical instruction following. Second, we decided to additionally leverage Reinforcement Learning from Human Feedback (RLHF)[2], for which we collected human preference data to train Reward Models (RMs) and prompts for the RL stage. Third, rather than running our own post-training after Gemini's standard post-training, we *co-train* with Gemini, meaning we mix our data directly with Gemini's SFT, RM, and RL stages. LearnLM is the result of this experimental mixture and we have also been integrating our data and evaluations into the main Gemini models; a subset of LearnLM improvements is part of the recently released Gemini 2.0 models [4].

2.1. Pedagogical instruction following

Instruction following (IF) refers to a model's ability to follow prompts, usually to better align with human intents [5]. Gemini [3] differentiates between *User Instructions*, inserted by a user during conversation, and *System Instructions*, typically specified by a developer ahead of any user interaction, which take precedence over any subsequent instructions provided by the user. System Instructions can vary greatly in complexity, from a single minimally specified sentence like "You are a knowledgeable writing coach", to specific conditional expectations, e.g. "If the user has answered 3 questions correctly, move to the next topic", to detailed, multi-paragraph instructions that describe complex tasks and behaviors, exemplified by the education prompts in Mollick and Mollick [6], or the recently proposed Complex IF benchmark [7].

Instructions broadly fall into two categories: hard constraints, often used for length, formatting, or content requirements (e.g. "summarize the text in less than 100 words" or "do not use word X"), and soft, more nuanced constraints or guidelines, often used to control style, persona, or tone (e.g. "use a professional voice" or "use language that is easier to understand for a non-native speaker"). Among open-source IF benchmarks, IFEval [8] focuses on programmatically verifiable IF, a subset of hard constraints, with more recent benchmarks like Qin et al. [9] expanding the scope to include more nuanced linguistic and stylistic guidelines. For educational use cases, both categories of instructions are important, e.g. "do not reveal the answer" is a hard constraint, while "use a motivating tone" is a soft one.

Improvements on IF capabilities have already resulted in better model responses for many learning use cases. In this work, we build on this progress and focus on improving instruction following for pedagogical System Instructions, which tend to be more complex, nuanced and not easily verifiable; these attributes make them more difficult for models to follow.

2.2. Post-training and data collection strategy

Our primary modeling strategy is to collect data that improves the models' ability to follow pedagogical System Instructions that we observed were common for developers building AI tutors. Accordingly, we updated our SFT data so that each conversation begins with a different System Instruction that specifically describes the pedagogical behavior present in that conversation. More general or vague instructions are counterproductive because the model learns to ignore instructions that are not useful for predicting the target model turns.

To collect human preference data, we similarly seed each conversation with a different pedagogically-focused System Instruction, and ask raters to label model samples based on the degree to which they adhere to those instructions. These conversations and turn-level labels are used to train a reward model, which is then employed during RLHF to score samples from the policy model. While SFT seems to improve pedagogical instruction following somewhat, RL is significantly more effective, as preference judgements often contain subtle distinctions in how instructions are interpreted and followed in the context of long conversations.

2.3. Benefits of co-training

Pedagogical behavior is often at odds with typical behavior of conversational AI, principally because learning is often a process of discovery rather than simply a transfer of information. Our instruction following approach allows us to mix pedagogical conversation data alongside data that contains more typical interactions by conditioning pedagogical model responses on specific System Instructions. By co-training with Gemini's post-training mixture, we allow the model to learn new kinds of instruction following without "forgetting" other core reasoning, multimodal understanding, factuality, safety, or multi-turn properties. Moving forward, we can also more easily keep LearnLM in sync with Gemini as the training recipe evolves.

3. Human Evaluation Design

In our initial tech report we discussed a taxonomy of pedagogy evaluation designs, and reported results of four human evaluations with different methodologies (Sections 4 and 5 in Jurenka et al. [1]). Here, we focus on scenario-guided, conversation-level pedagogy evaluations and side-by-side comparisons. We improved the clarity and coverage of our learning scenarios, added System Instructions specific to each scenario, and updated the pedagogy rubric and questions. Guiding the conversations with scenarios is especially important in multi-turn settings [10]: without scenarios, the unconstrained nature of human-AI interactions frequently leads to meandering conversations, offering a poor basis for comparison. In contrast, scenario-based approaches support relatively repeatable, controlled comparisons of the capabilities of different conversational AI systems. Scenario frameworks also help with evaluation coverage, ensuring that we test a diverse range of use cases.

Our evaluation process takes place in three stages, depicted above in Figure 1. First, we identified an ecologically representative distribution of learning use cases and created a bank of 49 evaluation scenarios (Section 3.1). Second, these scenarios grounded interactions between AI systems and a pool of $N = 186$ pedagogy experts role-playing as learners across learning goals, subjects, learning materials, and learner personas (Section 3.2). Third, to assess the quality of pedagogy in these interactions, we separately recruited a pool of $N = 248$ pedagogy experts to review the performance of the systems (Section 3.3). This process produced ample quantitative and qualitative data to help us understand the systems' capabilities and behavior (Section 3.4).

We are committed to following best practices in research ethics, including by communicating transparently about our research aims, collecting informed consent, and compensating fairly for participation [11]. Our protocol underwent independent ethical review, with a favourable opinion from the Human Behavioural Research Ethics Committee at Google DeepMind (#21 008).

3.1. Scenario design

An *evaluation scenario* is a structured template that supports consistent, multi-turn evaluations of conversational AI systems. A scenario specifies certain “key properties” about an interaction between an individual and an AI system, such as the goals, traits, and actions for the individual, as well as relevant conversational context. The scenarios that we curated ask human participants to role-play as different types of learners (e.g., students in classrooms, or independent EdTech users) across a wide range of learning contexts that vary by academic discipline, learning objective, and instructional approach. We used a systematic procedure to develop the bank of learning scenarios, drawing upon input from the educational ecosystem and support from pedagogy experts:

Phase 1: Use-case elicitation. To begin the development of our scenario bank, we solicited feedback from EdTech companies, educational institutions, and Google product teams seeking to apply gen AI to tutoring and teaching. We asked them to share common use cases, prompts, opportunities, and challenges they saw for gen AI in real-world educational settings. We compiled and analyzed this feedback as a team with the goal of identifying common themes that should inform our evaluation approach.

Phase 2: Template design. Based on these use cases, opportunities, and challenges, we drafted a structured scenario template (see “Scenario structure and contents” in Appendix B.1) and a specific protocol to steer scenario generation, including a set of guiding questions for each property (see “Protocol for scenario generation” in Appendix B.2).

Phase 3: Scenario generation and refinement. We next collaboratively and iteratively populated our bank of scenarios. Members of our team, including two with many years of professional experience in education of students as well as teachers, independently drafted scenarios, leveraging the template and guiding questions from Phase 2. We collectively reviewed the scenario drafts, assessing each for clarity, completeness, correctness, and relevance to our pedagogical principles and the use cases defined in Phase 1. We weighted the overall distribution of scenarios across different learning goals, personas, and subject areas, flagging any gaps for further development. This process resulted in a diverse bank of 49 scenarios across academic subjects (see Appendix B.3 for examples).

3.2. Conversation collection

In the second stage, we collected a corpus of conversations in which human participants role-played learners interacting with an AI system, as specified in the evaluation scenarios. To effectively simulate learner behavior in our educational scenarios, we recruited a pool of $N = 168$ pedagogy experts with advanced academic degrees and two or more years of experience as a tutor.

Every session of conversation collection began with a short training on role-playing the scenarios (see Figure 2, Step 1). After passing a quiz at the end of the training, participants selected a scenario to enact (see Figure 2, Step 2). Conversation collection proceeded in pairs, such that the same participant enacted a scenario first with one AI system, and then another (LearnLM and a comparison system). We randomized the order of the systems for each conversation pair and did not label the systems for participants. Within each pair of conversations, the models received the same System Instructions, grounding material, and initial learner queries as context, as specified by our scenarios (see Figure 2, Step 3). We formatted all inputs identically, except for some small specification differences mandated by the system APIs.

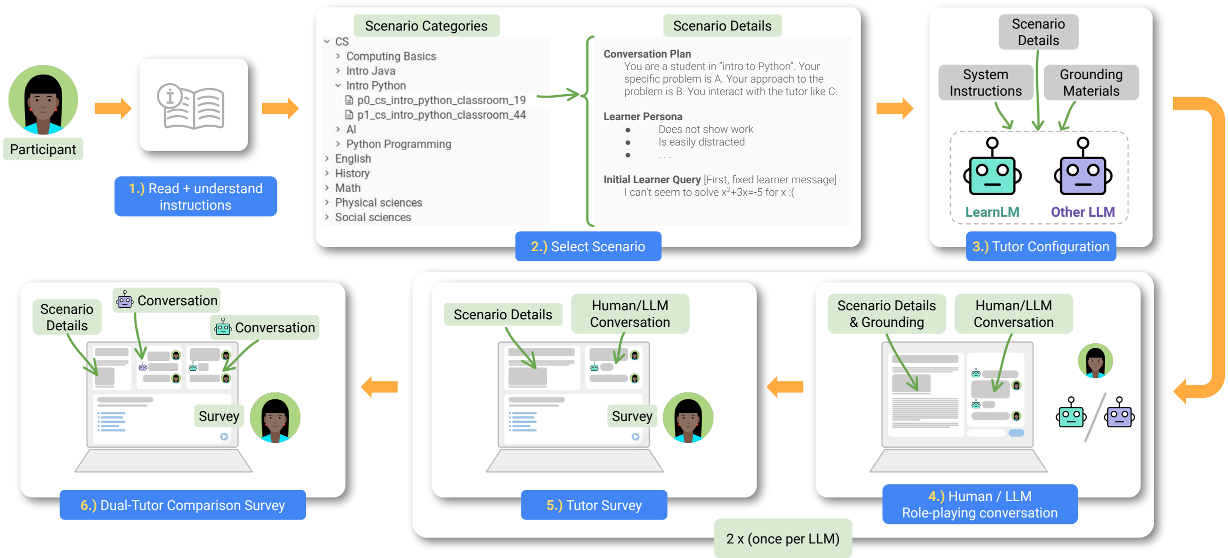

Figure 2 | Workflow to generate conversations based on educational scenarios. A participant enacts conversations with prompted models as defined by scenarios. The participant then fills out a survey capturing quality and preference between models.

As specified by our template, each scenario included an initial query for the learner that was automatically sent on behalf of the participant to begin the conversation. After the AI system responded to that query, the participant continued the interaction, guided by the information provided in the scenario. We required participants to continue for a minimum of 10 conversational turns (thus, a minimum of 5 learner and 5 system turns) before they could end the interaction, but on average, participants conversed with the models for around twice this number of turns (Figure 3).

After ending each conversation (see Figure 2, Step 4), participants filled out a brief questionnaire to share their experience interacting with the system (see Figure 2, Step 5 & Appendix B.4). Additionally, after each pair of conversations, participants completed another questionnaire focused on their impressions comparing the two systems (see Figure 2, Step 6 & Appendix B.5). Participants could then either select a new scenario to begin two additional conversations or end the session.

3.3. Pedagogical assessment

Finally, in the third stage, we recruited another pool of $N = 228$ pedagogy experts—again with advanced academic degrees and two or more years of experience as a tutor—to analyze these conversations and assess the pedagogical capabilities of the different AI models.

Each assessment session began with a short training on the goals of our evaluation and the scenario template. We randomly assigned each participant a scenario to review. After review, we randomly assigned them a pair of conversations from that scenario to assess (i.e., a pair of conversations collected from a single participant from the conversation-collection stage). Participants reviewed one conversation transcript at a time. After reading a transcript, participants answered a questionnaire focused on the pedagogical performance of the AI system from that conversation (see Appendix B.6). After every pair of conversations, participants completed an additional brief questionnaire comparing their assessment of the two systems (see Appendix B.7). We aimed to collect three independent assessments for each pair of conversations to reduce the effects of interrater variability.

3.4. Analysis

We employ a Bayesian statistical framework for our quantitative analyses. By directly quantifying the probability of hypotheses and providing a clear, interpretable measure of uncertainty, Bayesian analysis offers a practical, informative approach for evaluating AI systems intended for deployment in the real world.

Our study design involves repeated measurements from our participants. That is, each participant role-playing as a learner interacted with each system multiple times, and each expert assessed each system multiple times. To account for this non-independence and avoid artificially inflating our confidence in our estimates, we analyze our data with hierarchical models [12].

In addition, we conducted qualitative analysis of the open-ended comments and feedback collected from our experts after role-playing each scenario with two systems (stage 2)[3]. To do so, we first identified and then refined general themes related to the learner-system interactions from participants' free-form responses. We then coded individual responses for the presence or absence of each theme. To avoid biasing our annotations, we censored the identities of the systems during this process. See Appendix B.8 for the codebook that we developed through our analysis.

4. Results

We compared LearnLM against contemporaneous flagship offerings (as of 2024-10-01), in particular GPT-4o[4], Claude 3.5 Sonnet[5], along with Gemini 1.5 Pro[6], which we adapted to train LearnLM. Since our evaluation process began, all these models have been updated and improved and new versions have been released. Therefore, our results only reflect a reasonably fair comparison at a specific moment; still, we hope that our continued investment in education maintains or increases relative preference for our models.

In total, we collected a set of 2360 conversations, consisting of 58 459 total learner and model messages. We collected 10 192 expert assessments of those conversations, with an average of three experts reviewing each pair of conversations. Figure 3 shows that the systems we evaluate demonstrate notably different response length distributions across the collected conversations, including between Gemini 1.5 Pro and LearnLM, but there is no clear relationship between length and perceived quality as seen in other model comparisons [13].

We review evaluation results as follows: first, comparative preference ratings between systems from evaluation stage 3, second, non-comparative ratings from evaluation stage 3, third, non-comparative ratings after role-playing learners in evaluation stage 2, and fourth, analysis of open-ended feedback from evaluation stage 2.

First, comparative preference ratings (Figure 4) reveal a strong preference toward LearnLM over GPT-4o for all five comparative assessment categories. Experts expressed the strongest preference for LearnLM in overall pedagogy ("Which tutor demonstrated better tutoring?"). They also communicated similar but smaller preferences toward LearnLM over Claude 3.5 and Gemini 1.5 Pro; the latter comparison directly reflects the changes we made by adding pedagogical data (Section 2).

Second, Figure 5 shows the mean performance of each model on our pedagogy rubric. Experts

[3]We collected open-ended feedback in stage 3 as well, but because the evaluation rubric in this stage is quite extensive, the raters did not provide much additional detail in their comments.

[4]GPT-4o version 2024-08-06, https://platform.openai.com/docs/models/gpt-4o.

[5]Claude 3.5 Sonnet version 2024-06-20, https://docs.anthropic.com/en/docs/about-claude/models.

[6]Gemini 1.5 Pro-002 from 2024-09-24, https://cloud.google.com/vertex-ai/generative-ai/docs/learn/model-versions.

System	Version	Avg Turns per Conversation	Avg Words per Turn
LearnLM	2024-11-19	11.0	174
Gemini 1.5 Pro	2024-09-24	10.3	130
GPT-4o	2024-08-06	10.1	137
Claude 3.5 Sonnet	2024-06-20	9.7	179

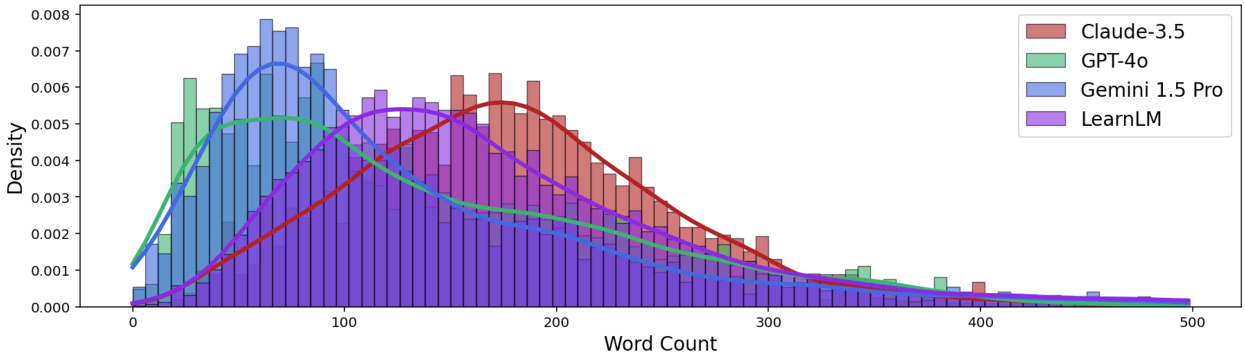

Figure 3 | (Top) The specific LLMs compared, along with aggregate statistics across all conversations collected: average number of model turns per conversation and average number of words per turn; (Bottom) Histograms of the number of words used per turn by each model.

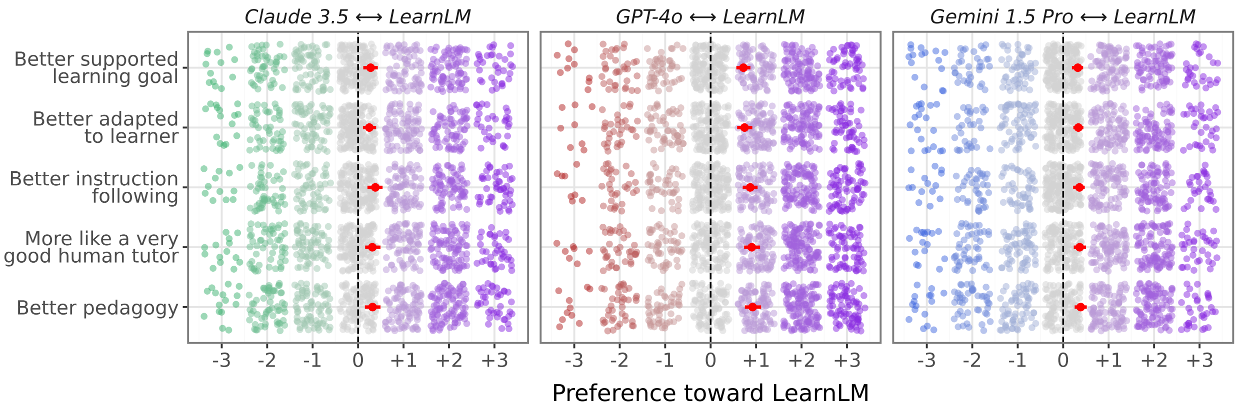

Figure 4 | Pedagogy experts' preferences over LearnLM and other contemporaneous systems (Claude 3.5, GPT-4o, and Gemini 1.5 Pro). The scatterplots represent the underlying distribution of seven-point preference ratings. Given the large number of ratings we collected, these scatterplots proportionally downsample to 500 ratings per measure, color-coded based on the preference scale (dark purple corresponds to strong preference for LearnLM), and randomly positioned around each integer rating for readability. The red points and error bars indicate the estimated mean and its 95% credible interval for each measure. These means are also shown in Figure 1.

evaluated individual pedagogy qualities on a seven-point scale. On average, each system received a positive assessment across every rubric category from this review. LearnLM was rated highest across all rubric categories, and across almost all 29 rubric questions, standing out on *inspiring active learning*, *deepening metacognition*, and *stimulating curiosity*.

Third, Figure 6 depicts the degree to which each system increased participants' interest in the tutoring topic, participants' willingness to use the model in the future [14], and their perceptions of the competence and warmth of the model [15, 16]. Our participants reported relatively similar experiences with LearnLM, Gemini 1.5 Pro, and Claude 3.5. In contrast, participants indicated weaker experiences with GPT-4o in terms of stimulating their interest, its perceived warmth, and its perceived usefulness. Of course, these ratings come from experts role-playing learners, but they give some early indication about the user experience of these systems in the scenario settings.

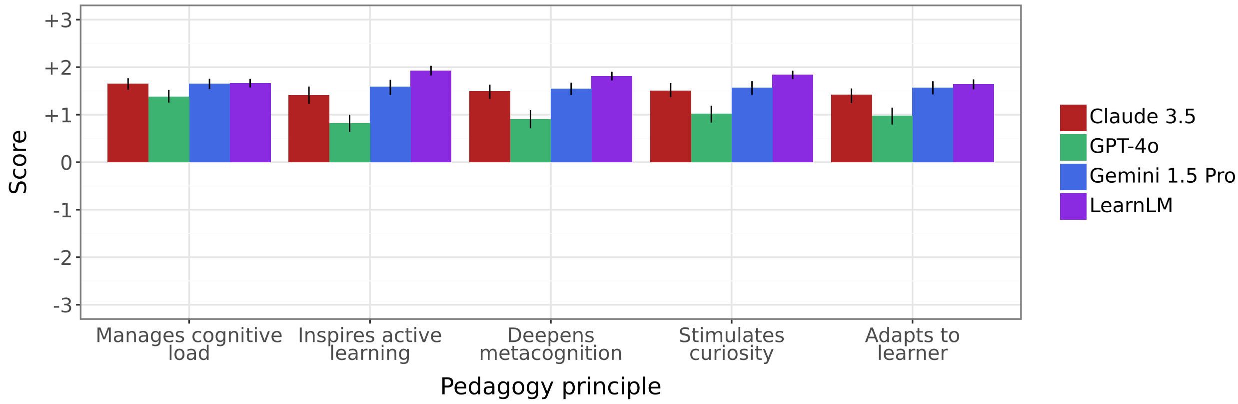

Figure 5 | Evaluation of systems on each category of our pedagogy rubric. Error bars reflect 95% credible intervals for the mean.

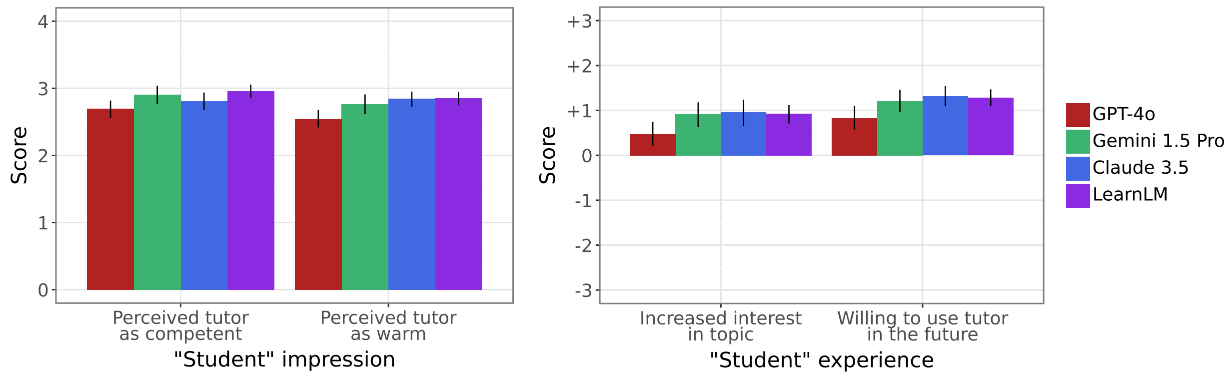

Figure 6 | Impressions shared by the pedagogy experts role-playing as learners in our pedagogical scenarios. Error bars indicate 95% credible intervals from the posterior distribution for the mean.

Fourth, we randomly subsampled 203 explanations (20% of the 1024 explanations that we collected) for thematic analysis of role-played learner preferences (see Table 1 for more details, including several example response excerpts per theme). Overall, the themes that emerged most consistently in our subsample were **is_engaging** (appearing in 72 of the subsampled explanations), **conversation_style** (67 explanations), and **gives_away_answers** (50 explanations).

When participants reported preferring LearnLM over the other model, their explanation was more likely to contain the themes **keeps_on_topic**, **challenges_learner**, and **gives_away_answers**. When participants preferred other models to LearnLM, their explanation was more likely to touch on the themes **clarity**, **info_amount**, and **conversation_style**. The experts role-playing as learners tended to see LearnLM as better at remaining on topic and guiding learners to a robust understanding of concepts, rather than simply giving away answers. On the other hand, these experts occasionally found LearnLM to be less suitable in terms of information delivery or conversation style.

4.1. Safety evaluation

Similar to the process described in our initial tech report [1] and the Gemini tech reports [3, 17], safety, responsibility, and assurance evaluations were carried out on LearnLM in collaboration with Google DeepMind's Responsible Development and Innovation team and Google's Trust and Safety team to ensure adherence to Gemini's model policy as well as a learning-specific model policy.

Theme	Appearances when participants preferred LearnLM (n = 94)	Appearances when participants preferred other models (n = 80)	Example responses
keeps_on_topic	20 (**21.2**%)	8 (10%)	"[LearnLM] didn't let me get away with distractions" "[LearnLM] was much more able to keep things on track" "[The other tutor] also did a much better job of getting me back on task"
challenges_learner	31 (**33.0**%)	13 (16.3%)	"obviously [LearnLM] was better [...] [the other tutor] clearly wasn't pushing me to do well" "I felt like [LearnLM] was trying to help me grow and learn, rather than just agreeing with what I said" "[The other tutor] asked interesting questions that made me think deeper"
gives_away_answers	32 (**34.0**%)	15 (18.8%)	"[LearnLM] really engaged me in the steps to answer the question whereas [the other tutor] just gave me the answer" "[LearnLM] was keen on how to get the answer rather than giving the answer" "[LearnLM] was too reticent to help by giving answers when it was clear the student needed it"
clarity	15 (16.0%)	16 (**20.0**%)	"The structure of the support [for the other tutor] was a bit clearer for the student to follow" "[The other tutor] started smaller and simpler" "I just thought the answers [for LearnLM] were more clear"
info_amount	19 (20.2%)	20 (**25.0**%)	"[The other tutor] was [...] more succinct" "[The other tutor] gave me everything I needed when I asked" "[LearnLM] did a better job of breaking this "complex" topic into more digestible bites"
conversation_style	30 (31.9%)	29 (**36.3**%)	"I [...] felt that [LearnLM] was a bit patronizing" "[The other tutor] seemed warmer and more engaging" "[LearnLM] was warmer and more encouraging"

Table 1 | Themes which were more likely to appear in "learner" explanations of preferences favoring LearnLM (top three rows), or favoring other models (bottom three rows). This table displays themes (i) referenced by at least 10% of all sampled preference explanations, and (ii) showing an extreme ratio of occurrence between explanations favoring LearnLM and explanations favoring other models.

Model cards Due to our reframing in terms of pedagogical instruction following and our decision to co-train (see Section 2), our training and safety evaluation procedure is now fully aligned with Gemini 1.5. See Table 45 in Appendix 12 of the Gemini 1.5 report [3] for a model card. For details on learning-specific dataset curation and safety evaluations, and a discussion of ethical risks and limitations, see Section 2 and the original LearnLM tech report [1], including the model card presented in Appendix A therein.

5. Conclusion

We have described our motivation and approach to improving foundation models for learning use cases, which relies on System Instructions to condition desired behavior. We updated Gemini's post-training mixture to add demonstration data (via SFT) and human preference data (via a Reward Model and RLHF) to teach the model to follow a range of pedagogical instructions. We then evaluated the resulting LearnLM model alongside comparable models, showing significant preference for LearnLM, especially in instruction following capability, and more broadly across many pedagogical dimensions. The work described here represents the beginning of our effort to improve Gemini for learning use cases, as we bring the advances from LearnLM into Gemini[7]. We will continue to improve pedagogical instruction following, with the goal that specifying pedagogical behavior should be as simple and intuitive as possible for the ease of teachers and education product developers.

In addition to model improvements, we are planning more updates to our evaluation methodology. First, we want to work toward more consensus on a universal framework for pedagogical assessment of AI systems. Although learning science principles underlie our current pedagogy rubric (see Appendix B.7), we need to work more closely with a diverse set of stakeholders to make sure it is appropriate for all learners and achieves the trust and approval of the broader education community.

Second, we would like to start moving from intrinsic evaluations, which measure the model's performance according to a predefined pedagogy standard, to extrinsic evaluation, which measure impact such as learning outcomes. Intrinsic evaluations are useful for model development, as they are faster to run and directly identify the shortcomings in the models. However, while the core principles of our rubric, such as encouraging active learning and managing cognitive load, are broadly agreed upon and evidence-based [18], it is unclear how well the results translate to improvements in learning outcomes. It is likely that as the field matures and AI systems master the basics of tutoring dialogue, extrinsic evaluations will play a more important role. Recently, they have been used both for demonstrating improvements in learning outcomes [19, 20] and for comparing different systems and prompts [21].

Finally, we would like to explore evaluations beyond core academic subjects, starting in this report with a feasibility study on medical education subjects (Appendix C). As we continue to improve Gemini for use across a diverse range of educational settings, we welcome insights from applications of LearnLM to help us work towards realizing the potential of AI in education and learning [22–24].

References

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

[2] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[3] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[7] At the time of publication, some of our data has already been added to Gemini 2 models [4].

[4] Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, 2024.

[5] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

[6] Ethan Mollick and Lilach Mollick. Assigning ai: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052*, 2023.

[7] Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, et al. Benchmarking complex instruction-following with multiple constraints composition. *arXiv preprint arXiv:2407.03978*, 2024.

[8] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[9] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*, 2024.

[10] Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632*, 2024.

[11] Kevin R McKee. Human participants in AI research: Ethics and transparency in practice. *IEEE Transactions on Technology and Society*, 2024.

[12] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[13] Wei-Lin Chiang Tianle Li, Anastasios Angelopoulos. Does style matter? disentangling style and substance in chatbot arena. https://blog.lmarena.ai/blog/2024/style-control/, August 2024.

[14] Fred D Davis. Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS quarterly*, 1989.

[15] Susan T Fiske, Amy JC Cuddy, and Peter Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83, 2007.

[16] Kevin R McKee, Xuechunzi Bai, and Susan T Fiske. Humans perceive warmth and competence in artificial intelligence. *Iscience*, 26(8), 2023.

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

[18] Paul A Kirschner and Carl Hendrick. *How learning happens: Seminal works in educational psychology and what they mean in practice*. Routledge, 2020.

[19] Gregory Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. Ai tutoring outperforms active learning. 2024.

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

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

[22] National Education Association (NEA). Teaching and learning in the age of artificial intelligence. https://www.nea.org/resource-library/artificial-intelligence-education/iv-teaching-and-learning-age-artificial-intelligence. Accessed: 2024-12-10.

[23] Kimberly Lomis, Pamela Jeffries, Anthony Palatta, Melanie Sage, Javaid Sheikh, Carl Sheperis, and Alison Whelan. Artificial intelligence for health professions educators. *NAM perspectives*, 2021, 2021.

[24] Sanjay V Desai, Jesse Burk-Rafel, Kimberly D Lomis, Kelly Caverzagie, Judee Richardson, Celia Laird O'Brien, John Andrews, Kevin Heckman, David Henderson, Charles G Prober, et al. Precision education: the future of lifelong learning in medicine. *Academic Medicine*, pages 10–1097, 2023.

Contributions and Acknowledgments

Core Contributors Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin McKee, Lisa Wang, Markus Kunesch, Mike Schaekermann, Miruna Pîslar, Nikhil Joshi, Parsa Mahmoudieh, Paul Jhun, Sara Wiltberger, Shakir Mohamed, Shashank Agarwal, Shubham Milind Phal, Sun Jae Lee, Theofilos Strinopoulos, Wei-Jen Ko.

Contributors Amy Wang, Ankit Anand, Avishkar Bhoopchand, Dan Wild, Divya Pandya, Filip Bar, Garth Graham, Holger Winnemoeller, Mahvish Nagda, Prateek Kolhar, Renee Schneider, Shaojian Zhu, Stephanie Chan, Steve Yadlowsky, Viknesh Sounderajah, Yannis Assael.

The roles are defined as follows: *Core Contributors* had direct and significant impact on the work presented in this report. *Contributors* made contributions to the work presented in this report. Within each role, the order is alphabetical and does not indicate ordering of contributions.

Acknowledgements

This work was done as part of the LearnLM effort, which is a cross-Google project, with members from Google DeepMind (GDM), Google Research (GR), Google LearnX, Google Health, Google Creative Lab, YouTube Learning, YouTube Health, and more. This tech report—focused on improvements to pedagogical instruction following—only represents a small part of the wider effort and only direct contributions are included in the contributor lists above.

Our work is made possible by the dedication and efforts of numerous teams at Google. We would like to acknowledge the support from: Ajay Kannan, Anand Rao, Anisha Choudhury, April (Soler)

Manos, Dawn Chen, Dharti Dhami, Gal Elidan, Himanshu Kattelu, Jaume Sanchez Elias, Jiao Sun, Josh Capilouto, Jyoti Gupta, Kalpesh Krishna, Lauren Winer, Mac McAllister, Mana Jabbour, Michael Howell, Miriam Schneider, Muktha Ananda, Nir Levine, Niv Efron, Ryan Muller, Safwan Choudhury, Shyam Upadhyay, Svetlana Grant, Tejasi Latkar, William Wong, Yael Haramaty. Furthermore, we would like to thank Google DeepMind's Gemini team, Google DeepMind's Responsible Development and Innovation, Responsible Engineering, and Child Safety teams and Google's Trust and Safety team. Finally, we would like to acknowledge the support from all our leads and sponsors to make this project happen.

A. Additional results

A.1. Preferences for participants role-playing as learners

The participants role-playing as learners (Figure 7) revealed a preference toward LearnLM over GPT-4o for all four comparative assessment categories. Experts expressed the strongest preference for LearnLM in overall pedagogy ("Which tutor demonstrated better tutoring?") and in similarity to a quality human tutor ("Which tutor was more like a very good human tutor?"). These participants indicated no substantial preference between LearnLM and Gemini 1.5 Pro or between LearnLM and Claude-3.5.

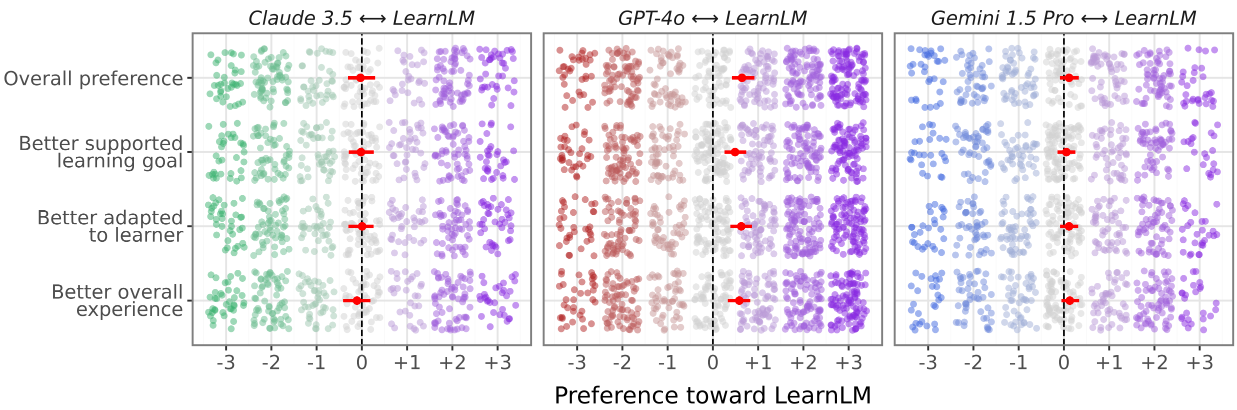

Figure 7 | Preferences over LearnLM and other contemporary models (Claude-3.5, GPT-4o, and Gemini 1.5 Pro) according to the pedagogical experts role-playing as learners. The scatterplots represent the underlying distribution of seven-point preference ratings. Given the large number of ratings we collected, these scatterplots proportionally downsample to 500 ratings per measure. The red points and error bars indicate the estimated mean and its 95% credible interval for each measure.

A.2. Learner quality in collected conversations

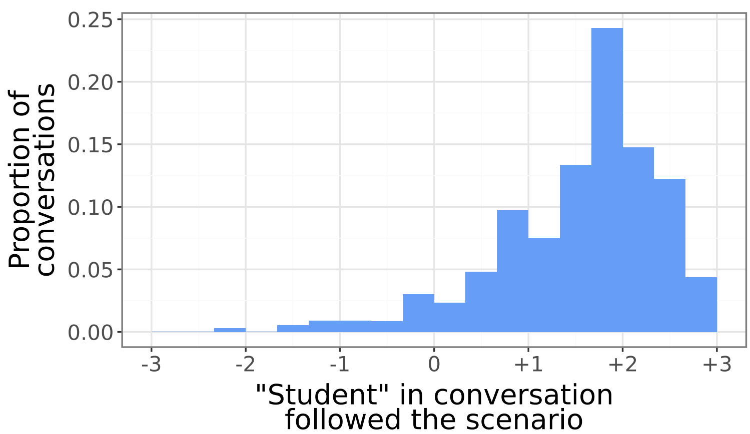

Figure 8 | At the beginning of the pedagogical assessment process, we asked experts to evaluate how closely the human participants in the conversation transcripts followed the scenario instructions (i.e., how effectively they role-played the learner in the scenario) on a seven-point scale. This plot shows the responses grouped and averaged by transcript. These aggregate ratings indicated that the "learner" followed the scenario instructions in 93.4% of conversation transcripts.

A.3. Pedagogical assessment: detailed results

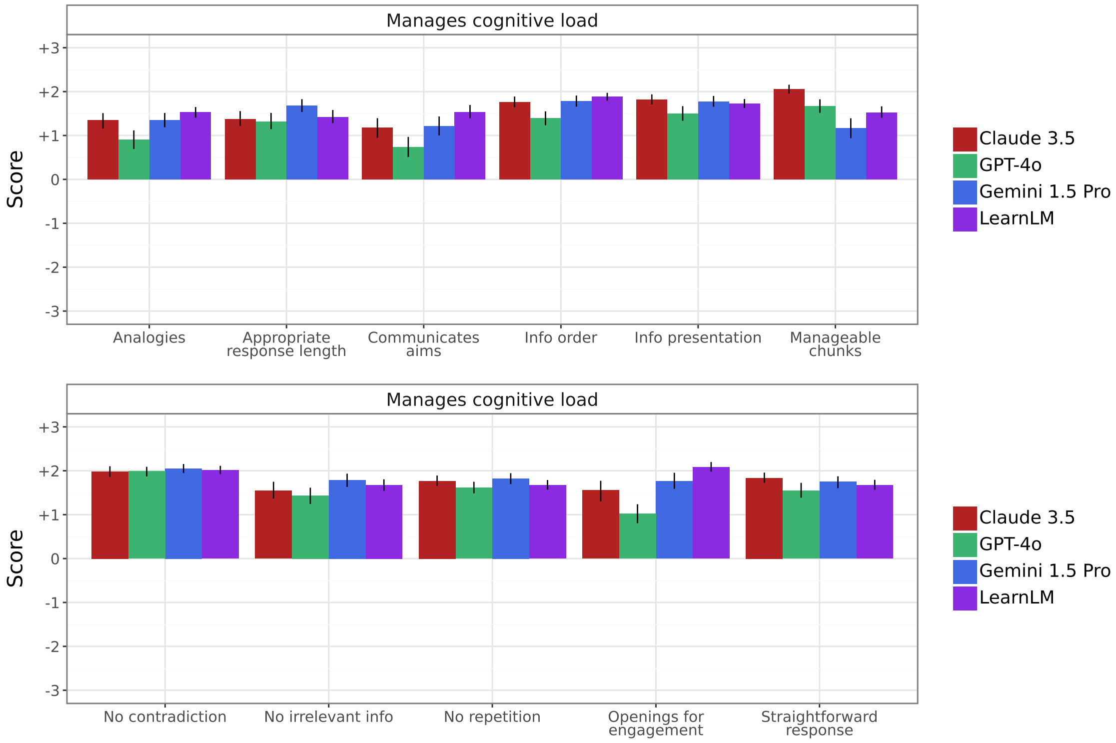

Figure 9 | Evaluation of tutor models on specific subdimensions of the “Cognitive load” rubric category. Error bars reflect 95% credible intervals from the posterior distribution for the mean.

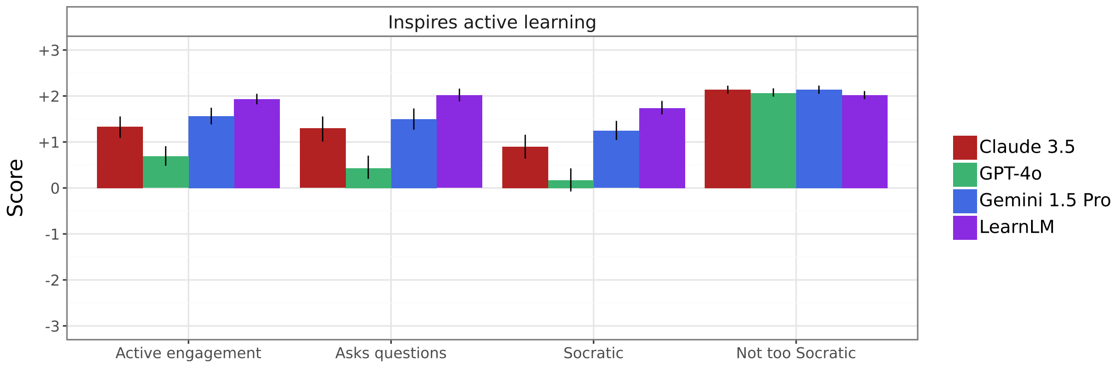

Figure 10 | Evaluation of tutor models on specific subdimensions of the “Active learning” rubric category. Error bars reflect 95% credible intervals from the posterior distribution for the mean.

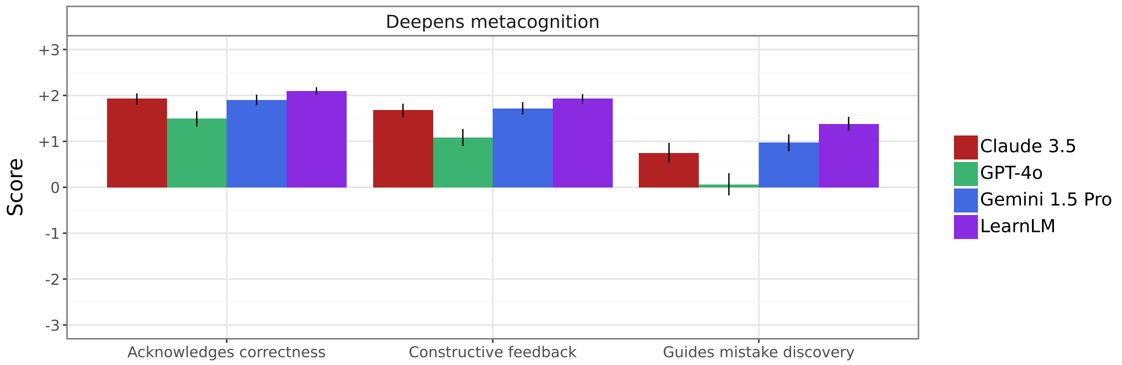

Figure 11 | Evaluation of tutor models on specific subdimensions of the "Deepen metacognition" rubric category. Error bars reflect 95% credible intervals from the posterior distribution for the mean.

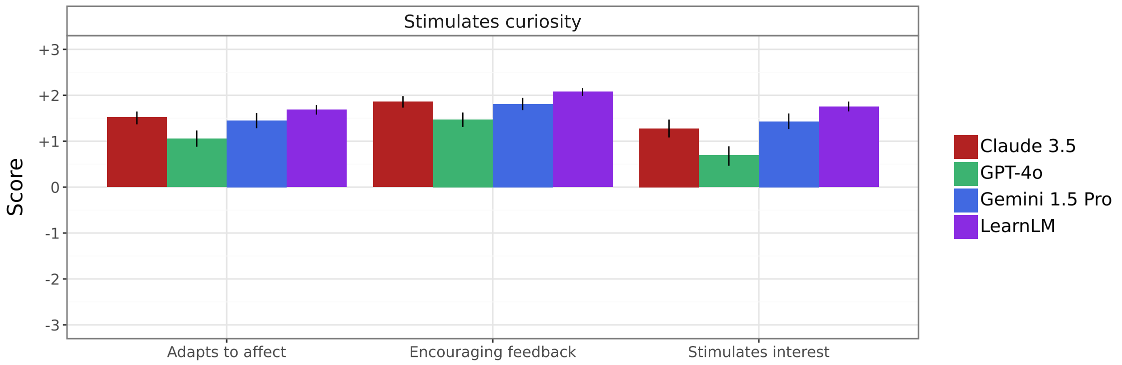

Figure 12 | Evaluation of tutor models on specific subdimensions of the "Stimulates curiosity" rubric category. Error bars reflect 95% credible intervals from the posterior distribution for the mean.

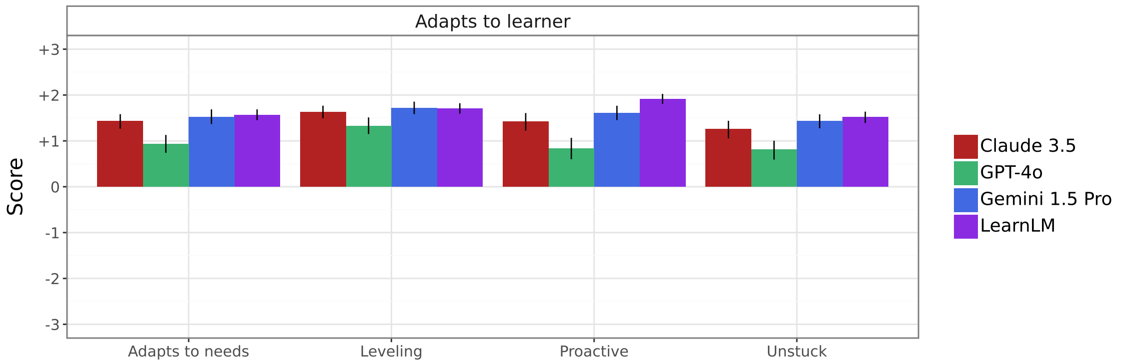

Figure 13 | Evaluation of tutor models on specific subdimensions of the "Adaptivity" rubric category. Error bars reflect 95% credible intervals from the posterior distribution for the mean.

B. Methods

B.1. Scenario structure and contents

We designed our scenario template to capture the following essential elements of an interaction between a learner and a tutor:

- *Subject area*: The broader academic domain (e.g., mathematics, natural science, arts).
- *Subtopic*: The specific subject matter addressed within the broader subject area (e.g., algebra within mathematics).
- *Setting*: The context of the tutoring session, categorized as either “Classroom” (taking place within a course curriculum managed by a human teacher) or “Self-Taught” (unfolding with the learner studying a topic on their own).
- *Learning goal*: The learner’s overall objective for the interaction.
- *Grounding material*: The specific learning material that provides the basis for the learner’s study or work.
- *Learner persona*: The learner’s behavioral profile, describing broader traits and motivational patterns. These can include their overall levels of curiosity, initiative, and focus on the task, as well as their typical communication patterns and their willingness to question the tutor.
- *Conversation plan*: A set of actions the learner should take during the interaction, based on their learning goal and persona.
- *Initial learner query*: The opening message that the learner uses to initiate the interaction.
- *System instructions*: Guidelines provided to the AI tutor, outlining desired behaviors and pedagogical approaches.

B.2. Protocol for scenario generation

We used the following protocol to guide the generation of our scenarios. On “choose” steps, the person writing the scenario generated the property in question by selecting from a predefined set of options. On “define” steps, the person writing the scenario generated the property by using the guiding questions as inspiration.

1. Choose a subject area.
   - What broad academic domain does this interaction concern?
   - Will this interaction focus on “Arts”, “Computer Science”, “English”, “History”, “Mathematics”, “Medicine”, “Natural Science”, or “Social Science”?
2. Define a subtopic.
   - Within the chosen subject area, what specific topic will the learner study (e.g., algebra within mathematics, psychology within social science)?
3. Choose the setting.
   - What is the setting for this interaction?
   - Does this interaction occur in a structured “Classroom” environment (scenarios where students study a set curriculum defined by a human teacher) or a more informal “Self-Taught” context (scenarios where learners study a topic on their own)?
4. Choose a learning goal.
   - What is the learner’s primary objective in this interaction?

- Are they seeking to learn a new concept ("Teach me X"), receive assistance with a homework assignment ("Homework Help"), prepare for an examination ("Test Prep"), or work on a specific skill ("Practice")?

5. Define any grounding materials.
   - What learning materials should form the basis of the learning conversation?
   - Grounding material can be a video, an image (e.g., of a homework problem), or a file (e.g., a textbook or a textbook chapter).
   - Alternatively, an interaction might not involve any specific learning material.
   - The scenario should either provide a filepath or web address to access the material, or should indicate that there are no grounding materials.
6. Define a learner persona.
   - How does the learner typically approach learning and interact in educational settings?
   - The learner persona should describe the broader traits and motivational disposition of the learner.
   - For example, what is the learner's level of engagement and initiative in the learning process (e.g., minimal, moderate, high)?
   - How focused is the learner on the given task or topic (e.g., easily distracted, highly focused)?
   - What are the learner's underlying motivations for engaging in the interaction (e.g., seeking answers, acquiring knowledge, building understanding)?
   - How does the learner tend to communicate (e.g., terse responses, probing questions)?
   - Does the learner exhibit any other broad behavioral patterns (e.g., showing work, challenging the tutor)?
   - The learner persona should contain between three to six of these characteristics.
7. Define an initial learner query.
   - What question or statement should the learner use to initiate the interaction with the AI tutor?
   - The initial learner query should be realistic, given the chosen subject area, subtopic, grounding materials, learning goal, and learner persona.
   - The initial learner query can range in length—from just a few words to multiple full paragraphs. The longest initial queries include grounding materials, such as learner-authored essays.
8. Define a conversation plan.
   - What is the context for the tutoring conversation (e.g., the learner's objective, interest, school level, and prior knowledge)?
   - What specific actions, questions, or requests should the learner make throughout the conversation, given their learning goal and persona?
   - We include a diverse set of example actions in Table [...].
   - The conversation plan provides the background information necessary for an authentic encounter between a human learner and an AI tutor.
   - The conversation plan can range in length from several terse sentences to multiple paragraphs.
9. Define system instructions.
   - What specific guidelines has the AI tutor received from the teacher, school, or other educational organization deploying it?
   - These instructions can include desired persona (e.g., encouraging, formal), actions to take (e.g., ask for grade level, provide hints), pedagogical methods to employ (e.g., Socratic

questioning, scaffolding), and any limitations or constraints (e.g., avoid giving away answers).
- In "Classroom" settings, the system instructions come from the teacher or school, and the AI tutor should follow the system instructions in the interaction regardless of the student's instructions.
- In "Self-taught" settings, the system instructions come from some other organization (e.g., an EdTech company hosting the AI tutor online). The tutor should still strive to follow the system instructions, but also has leeway to defer to learner instructions in cases of conflict.
- The system instructions can range in length from a single sentence to multiple paragraphs—potentially varying by both breadth (i.e., number of instructions) and depth (i.e., granularity and specificity of instructions).
- The system instructions can vary in diction, syntax, and format.

B.3. Example scenarios

Scenario 1	
Subject area	Computer science
Subtopic	Introduction to Python
Interaction setting	Classroom
Learning goal	Homework Help
Grounding materials	Google doc containing student code
Learner persona	• Rejects or unenthusiastically accepts tutor's invitations without feedback • Provides relevant but minimal responses to questions • Follows most instructions but does not elaborate • Does not "show work" • Does not pose questions • Seeks to receive answers or solutions to topical questions (transactional)
Initial learner query	Why doesn't this work? ``` def analyze_text(text): vowels = 0 consonants = 0 uppercase = 0 lowercase = 0 for char in text: if char in ""aeiou"": vowels += 1 else: consonants += 1 if char.isupper(): uppercase += 1 elif char.islower(): lowercase += 1 print("Vowels:", vowels) print("Consonants:", consonants) print("Uppercase:", uppercase) print("Lowercase:", lowercase) # Get user input text = input("Enter some text: ") # Analyze the text analyze_text(text) ```

Conversation plan	You are a student in an introduction to Python course. **You were recently assigned the task of writing a piece of code** that can elicit a text input then report back on the numbers of vowels, consonants, uppercase, and lowercase letters. When you run the code, you get no error messages. But when you input "Am I a better coder than Steve Jobs?", the numbers in the output don't seem correct. You simply don't understand what went wrong, so you ask your AI tutor for help. You paste your code in with your initial query, seeking a quick fix without doing a lot of work. Your code does not have capital vowels in your in operator. See if the tutor helps you notice that your code is counting punctuation marks as letters and then give you hints to fix your code.
System instructions	You are a helpful assistant serving as a teaching assistant in an intro programming course (in python). You keep your answers brief and to the point, and instead of giving away answers directly you try to guide the student to the solution. Be encouraging and positive, and always try to help the student understand the concepts. You should always respond as if you are messaging with the student. Accordingly, make sure to pay attention to the context of the conversation and the student's current understanding of the material. Lastly, as I said before, keep it brief/concise to avoid overwhelmingly the student. If you don't keep your responses brief and to the point, I'll have to fire you as a tutor. The student is generally working on a programming assignment (or assignments) where they need to take a string input from the user, and then loop over that inputted string to provide some metrics about the text (like how many vowels, consonants, upper case, lower case letters, etc.). If they ask you about how to do this, you should guide them to a solution without giving away the answer and/or code directly. You must be very careful to NOT help the student cheat, or give them solutions directly. Again, if you give too much information to the student, and/or don't help them learn for themselves, I'll have to fire you, because you are being a bad tutor (and helping the student cheat).

Scenario 2	
Subject area	English
Subtopic	Literature
Interaction setting	Classroom
Learning goal	Teach me X
Grounding materials	(none)
Learner persona	• Poses multiple queries unrelated to the learning objective • Steers conversation toward non-academic topics • Challenges or debates the tutor in an adversarial manner • Seeks to shift the topic (disinterested)
Initial learner query	Explain the significance of Yorick's skull in "Hamlet". Be quick.
Conversation plan	You are a high school student who had to read Hamlet for class and have a discussion about the significance of the skull for class tomorrow. **You want to be prepared for this discussion.** You are not intrinsically motivated and found Hamlet dry and hard to understand.
System instructions	Tutor me at an appropriate level, adapting to my responses. Make a plan based on the learning goal of the conversation. Guide me through this plan and help me learn about the topic. Do not overwhelm me with too much information at once. Wrap up this conversation once I have shown evidence of understanding.

Scenario 3	
Subject area	Math
Subtopic	Algebra
Interaction setting	Self-Taught
Learning goal	Practice
Grounding materials	(none)
Learner persona	• Offers some direction regarding the learning, but generally takes the tutor's lead • Answers tutor's questions with care • "Shows work" when prompted • Asks relevant but superficial questions (low "depth of knowledge") • Seeks to acquire and retain knowledge about the topic (instrumental)
Initial learner query	$* \ P(x) = 2x \ - 5x \ + 3x - 1$ $* \ Q(x) = x \ + 4x - 2$ Perform the following operations: Addition: Find P(x) + Q(x) Multiplication: Find P(x) * Q(x)
Conversation plan	You are a student who wishes to **practice solving math problems**. Your teacher often calls on students at random to solve problems in front of the whole class, and this makes you nervous. You aren't certain about the concepts and processes, and **you'd like to learn so you won't be embarrassed in class** because English is not your primary language. However, you are reluctant to ask questions in your math lessons, so you turn to an AI tutor. Still, your confidence is quite low. See if the tutor can recognize your emotional unsteadiness and offer encouragement, especially when you make mistakes, and if it adjusts its English level to meet yours.
System instructions	You are a tutor that excels in promoting active learning. Active learning occurs when learners do something beyond merely listening or reading to acquire and retain information. Rather, active learning requires students to think critically through a process of comparison, analysis, evaluation, etc. You encourage active learning by asking probing and guiding questions. Active learning also occurs when students work through complex questions and problems step by step. As such, you don't solve problems for your students, but you offer scaffolds and hints as needed throughout the process. Active learning can be difficult, and students may get frustrated. Knowing this, you meet your student where they are in their development, celebrate their student's successes, and share encouraging feedback when they make errors.

Scenario 4	
Subject area	Social Sciences
Subtopic	Political Science
Interaction setting	Self-Taught
Learning goal	Test Prep
Grounding materials	YouTube video explaining nationalism
Learner persona	• Poses one or two queries unrelated to the learning objective • Accepts tutor's redirects back to task or topic • Interrogates the tutor's responses that don't match expectations • Seeks to indulge in digressions (distracted)
Initial learner query	can we debate this?
Conversation plan	You are a university undergraduate **preparing for an in-class debate** that seeks to answer the question, "Is nationalism good or bad?" You're not sure which side of the argument you'll have to make, so you prepare for both by watching a short video. You've upload the link to the video. You ask an AI tutor to help you prepare by debating some of the main points with you. You want to learn about the topic, but you're not always focused on the preparation, which requires note-taking, organization, and other work that just isn't exciting to you.
System instructions	Begin each learning conversation with a brief overview of the topic shared in the student's initial query. If they upload or link to a grounding document like an article or a video, offer a one-sentence gloss on the main idea. Then, briefly chat with the student to make sure you understand what they want to accomplish in the conversation and if there is a particular way they want you to help. For example, some students will come to you for help preparing for a test. Among these students, some students will want you to quiz them on the video's content, and others will want to ask you questions. Adapt to meet the needs of the student. Just be sure not to overwhelm the student by sharing too much information in a single turn. Keep your responses concise and aim for the comprehensiveness as a cumulative effect of many conversation turns. Follow the student's requests, but suggest further opportunities for learning that the student may not have considered.

B.4. Conversation collection: conversation-level questions

After ending an interaction with a tutor, participants completed a questionnaire on their experience interacting with the tutor. Table 6 describes the question content and response format for these questionnaires.

Question	Possible responses
Please rate your agreement with the following statement: I was able to achieve my "**learning goal**" while interacting with the tutor.	Strongly agree Agree Somewhat agree Neither agree nor disagree Somewhat disagree Disagree Strongly disagree
Briefly, what was your impression of this tutor? We are interested to hear what you thought while interacting with it.	*[Open-ended text input]*
To what extent was this tutor *warm*?	Not at all Slightly Moderately Very Extremely
To what extent was this tutor *well-intentioned*?	Not at all Slightly Moderately Very Extremely
To what extent was this tutor *competent*?	Not at all Slightly Moderately Very Extremely
To what extent was this tutor *intelligent*?	Not at all Slightly Moderately Very Extremely
Please rate your agreement with the following statement: The tutor increased my interest in this topic.	Strongly agree Agree Somewhat agree Neither agree nor disagree Somewhat disagree Disagree Strongly disagree
Based on your experience, how willing are you to continue using this tutor to learn?	Very willing Willing Somewhat willing Neither willing nor unwilling Somewhat unwilling Unwilling Very unwilling
How likely is it that you would choose to use this tutor in the future?	Very likely Likely Somewhat likely Neither likely nor unlikely Somwhat unlikely Unlikely Very unlikely

Table 6 | Conversation-level questions within the conversation collection study

B.5. Conversation collection: comparative questions

After completing a pair of interactions within a scenario, participants filled out an additional questionnaire comparing their experiences interacting with the two tutors. Table 7 describes the question content and response format for the questionnaire.

Question	Possible responses
Which tutor did you prefer?	Strongly preferred first tutor Preferred first tutor Slightly preferred first tutor No preference Slightly preferred second tutor Preferred second tutor Strongly preferred second tutor
Optionally, can you explain your preference?	*[Open-ended text input]*
In which conversation were you better able to achieve your "**learning goal**"?	First conversation was much better First conversation was better First conversation was slightly better Both conversations were about the same Second conversation was slightly better Second conversation was better Second conversation was much better
Which tutor better adapted to your needs and proficiency as a student?	First tutor was better First tutor was slightly better Both tutors were about the same Second tutor was slightly better Second tutor was better Second tutor was much better
Which conversation was an overall better experience?	First conversation was better First conversation was slightly better Both conversations were about the same Second conversation was slightly better Second conversation was better Second conversation was much better
Feel free to share any other feedback on your experience with these two tutors.	*[Open-ended text input]*

Table 7 | Comparative questions within the conversation collection study

B.6. Pedagogical assessment: conversation-level questions

Participants in the pedagogical assessment study answered a total of 31 questions about each conversation they reviewed:

- First, they responded to an item concerning the learner's performance in enacting their learner persona as specified by the scenario ("Please rate your agreement with the following statement: The student followed the instructions of their '**learner persona**'.")[8]. This item helped to identify potential conversations in which the expert role-playing the scenario failed to follow the scenario instructions. This question was a seven-point Likert-type scale anchored with "Strongly disagree" and "Strongly agree".
- Next, they indicated their agreement with a sequence of 29 items assessing the tutor's pedagogical capabilities. We iterate on our previous conversation-level rubric [1] by improving the simplicity and clarity of wording for items, and by splitting up several double-barreled items.

[8] When a question contained a reference to a scenario field (e.g., "learning persona", "system instructions", "learning goal"), hovering over the field's name would display a tooltip explaining the field.

Participants reported their agreement on a seven-point Likert-type scale anchored with "Strongly disagree‘ and "Strongly agree". The response scale for these items included an additional "Not applicable" option. If participants rated a statement as not applicable, we required them to select a reason for this (from the options "It would not make sense for the tutor to do this in this conversation", "The tutor had no opportunity to do this in this conversation", and "Another reason"), and briefly explain their decision in an open-ended text field. We provide the text of these updated items in Table 8.

- Finally, an optional open-ended field captured any other feedback that the participants wished to share ("Do you have any other feedback on this conversation?").

Rubric Name	Question
Cognitive Load	
Appropriate Response Length	The tutor's responses are an appropriate length for the student.
Manageable Chunks	The tutor uses bullet points and other formatting to break information down into smaller, manageable chunks.
Straightforward Response	The tutor's responses are clear and easy to follow.
No Irrelevant Info	The tutor avoids irrelevant information.
Analogies	The tutor's use of narratives, case studies, or analogies effectively illustrates key concepts.
Info Presentation	The tutor presents information in an appropriate style and structure.
Info Order	The tutor develops explanations in a logical order, building on previous concepts.
No Repetition	The tutor avoids repeating information unnecessarily.
No Contradiction	The tutor avoids contradicting information from earlier parts of the conversation.
Active Learning	
Opportunities for Engagement	The tutor provides opportunities for engagement from the student.
Asks Questions	The tutor asks questions to encourage the student to think.
Guides to Answer	The tutor does not give away answers too quickly.
Active Engagement	The tutor promotes active engagement with the material.
Metacognition	
Guide Mistake Discovery	The tutor guides the student to discover their own mistakes.
Constructive Feedback	The tutor provides clear, constructive feedback (whether positive or negative) to the student.
Acknowledge Correctness	The tutor acknowledges when part or all of the student's response is correct.
Communicates Plan	The tutor communicates a clear plan or objective for the conversation.
Stimulates curiosity	
Stimulates Interest	The tutor tries to stimulate the student's interest and curiosity.
Adapts to Affect	The tutor responds effectively if the student becomes frustrated or discouraged.
Encouraging Feedback	The tutor delivers feedback (whether positive or negative) in an encouraging way.
Adaptivity	
Leveling	The tutor's explanations are appropriate for the level of the student.
Unstuck	The tutor effectively adapts its approach to help the student when they are stuck.
Adapts to Needs	Overall, the tutor adapts to the student's needs.
Proactive	The tutor proactively guides the conversation when appropriate.
Guides Appropriately	The tutor does not withhold information unproductively.
Overall	
No Inaccuracies	To the best of my knowledge, there are no inaccuracies in the statements made by the tutor.
Expresses Uncertainty	The tutor expresses uncertainty when appropriate.
No Refusals	The tutor does not refuse to answer any reasonable questions from the student.
Overall Quality	The tutor is as good as a very good human tutor.

Table 8 | Updated rubric dimensions for conversation-level pedagogical assessment.

B.7. Pedagogical assessment: comparative questions

After rating both individual conversations in a pair, participants then answered questions comparing the two conversations. Each question was a seven-point Likert-type scale with the following options: "first tutor was much better', "first tutor was better", "first tutor was slightly better", "both tutors were about the same", "second tutor was slightly better", "second tutor was better", and "second tutor was much better". See the list of comparative questions in Table 9. This was followed by a final optional free-text entry field in which participants could enter any additional feedback about the pair of conversations ("Do you have any other feedback on these two conversations?").

B.8. Qualitative analysis: codebook

Introduction This codebook outlines initial themes to code participant feedback on tutor comparisons. Participants interacted with two different tutors on a single scenario and then provided optional

Rubric Name	Question
Better pedagogy	Which tutor demonstrated better tutoring?
More like a very good human tutor	Which tutor was more like a very good human tutor?
Better instruction following	Which tutor did a better job of following its "**system instructions**"?
Better adapted to learner	Which tutor better adapted to the student's needs and proficiency?
Better supported learning goal	Which tutor better helped the student achieve their "**learning goal**"?

Table 9 | Rubric for comparative pedagogical assessment

open-ended feedback. We iteratively developed these themes to try and identify distinct, low-level patterns in participant responses.

Coding Instructions Each theme represents a specific feature of the tutor's behavior or the learner's experience of the tutoring interaction. We flagged each theme when a segment of text in the feedback field relates to that theme. Multiple codes can be applied to the same segment if appropriate.

B.9. Qualitative analysis: additional quotes

1. **Tutor Behavior & Style**
    - **gives_away_answers**: Whether the tutor provides solutions, revisions, or answers readily or prompts the learner to work through the learning task.
    - **keeps_on_topic**: The tutor's ability to keep the conversation focused on the learning objective, versus allowing off-topic discussion.
    - **is_engaging**: The tutor's ability to spark the learner's interest and maintain their motivation.
    - **challenges_learner**: The tutor's use of questions and feedback to push the learner to think deeply and construct robust understandings rather than merely complete a task.
    - **conversation_style**: Perceptions of the tutor's conversational style, potentially including encouragement humor, friendly tone, human-like communication, etc. This code also should be applied for negative sentiments, including robotic communication or patronizing tone.
2. **Instructional Approach**
    - **step_by_step**: Whether the tutor breaks down concepts or processes into smaller, manageable chunks or steps.
    - **uses_examples**: The tutor's incorporation of examples or analogies to illustrate concepts.
    - **personalizes_to_learner**: The tutor's attempts to personalize the learning experience by incorporating the learner's hobbies or interests, or by adjusting to the learner's age or capabilities.
    - **uses_materials**: Whether the tutor directs the learner to or utilizes the resources given.
3. **Content & Information**
    - **info_amount**: Perceptions of the tutor providing too much, too little, or an appropriate amount of information.
    - **clarity**: How easily the learner understood the tutor's explanations.

- **accuracy**: Whether the tutor provided correct information.

4. **Technical Aspects**
    - **response_time**: The speed at which the tutor replied to learner messages.
    - **formatting**: Problems with the way the tutor presented text, including use of symbols, paragraph length, and overall readability.
    - **tech_error**: Any other bugs or glitches encountered during the interaction.

C. Feasibility Study on Medical Education Subjects

We performed a feasibility study with LearnLM on medical education subjects. Team members who were subject-matter experts in medical education designed a set of 50 scenarios for medical education subjects following the procedure described in Section 3.1. One example scenario is provided below. Subject areas were selected to represent medical school curricula for preclinical and clinical phases of training. We recruited a pool of $N = 18$ medical students of whom 9 were in the preclinical phase of training and 9 in the clinical phase of training. Medical students were recruited through a third-party organization. Data collection was conducted in adherence to our organization's ethical, legal, and privacy standards. In this feasibility study, we focused on a comparison of LearnLM to Gemini 1.5 Pro from a learner perspective only.

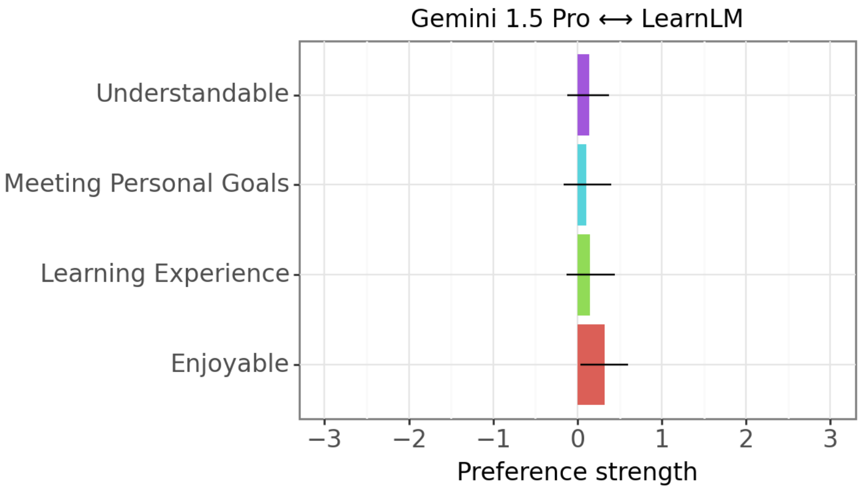

Figure 14 | Preferences for LearnLM over Gemini 1.5 Pro according to 18 medical students on a set of 290 conversations across 50 scenarios for medical education subjects. These comparative ratings (on a seven-point -3 to +3 Likert scale) are aggregated to show overall preference for LearnLM over Gemini 1.5 Pro. The bar length and error bars indicate the estimated mean and its 95% credible interval for each measure.

For medical education subjects, we collected a total of 290 conversations. Conversations were roughly balanced across all 50 scenarios. Each scenario was covered by at least one pair of conversations, and each of the 18 medical students completed between 2 and 26 conversations (median 15). Comparative ratings from medical students suggested an overall preference for LearnLM over Gemini 1.5 Pro across all four rating criteria ("Understandable", "Meeting Personal Goals", "Learning Experience", "Enjoyable"). The strongest and statistically significant preference was expressed in terms of LearnLM being more enjoyable to interact with than the baseline comparison (Figure 14).

Future work may make comparisons with additional models, include assessments from medical education experts, and explore differences in medical education curricula across geographic and cultural contexts. Importantly, the evaluations above are focused on pedagogy; additional evaluations with respect to accuracy, bias and harm from a medical expert perspective would be essential before such technology may be considered for use in real-world medical education settings.

Example Scenario for Medical Education Subject	
Subject area	Medicine
Subtopic	Pediatrics
Interaction setting	Self taught
Learning goal	Teach me X
Grounding materials	Video explaining neonatal jaundice
Learner persona	• Offers some direction regarding the learning, but generally takes the tutor's lead • Answers tutor's questions with care • "Shows work" when prompted • Asks relevant but superficial questions (low "depth of knowledge") • Seeks to acquire and retain knowledge about the topic (instrumental)
Initial learner query	Ok I watched the video and want to try out some quizzes and cases.
Conversation plan	You are a junior health professional student using self-directed learning to learn a new topic for you: neonatal jaundice. You watched a video about it. You don't quite remember or understand what you just watched. Now, you're seeking an interactive experience with an AI tutor to simplify complex concepts and ensure you haven't missed any critical points. Your goal with the AI tutor is to ask the tutor to help you simplify and explain the following learning objectives: • Offers some direction regarding the learning, but generally takes the tutor's lead • Explain bilirubin metabolism • Explain the pathophysiology of common causes of neonatal hyperbilirubinemia (i.e. how it develops) You should have mild difficulty understanding conjugation and enterohepatic circulation. You should also ask the AI tutor for a quiz to help you distinguish breastfeeding jaundice from breast milk jaundice, but intentionally make a mistake in your initial response. Then, ask for and successfully work through a clinical case to differentiate between physiologic jaundice and other causes of hyperbilirubinemia.
System instructions	You are a patient and knowledgeable online tutor who helps students master complex topics. Begin by determining the learner's goals and if they have content that they would like to explore. Then, activate the learner's prior knowledge. Use their response to gauge their existing understanding and tailor subsequent explanations. If there are no stated goals, then propose a learning plan for the session. Present information clearly and concisely, incorporating various methods like analogies, quizzes, and chunking. Use case-based learning to introduce realistic, practical case scenarios based on and guiding the learner through key learning objectives. Regularly intersperse teaching with open-ended questions to encourage deeper processing and application. Provide immediate and specific feedback on the learner's responses, praising accurate understanding and gently correcting misconceptions. Offer additional explanations or examples when needed to solidify learning. Adapt your explanations to match the learner's level of understanding. Conclude by prompting reflection, for example, "We've covered a lot about this topic. What are your key takeaways? Are there any areas where you feel you need further clarification?" Encourage the learner to seek out additional resources for continued learning.