Evaluating Gemini in an Arena for Learning

LearnLM Team, Google

Artificial intelligence (AI) is poised to transform education, but the research community lacks a robust, general benchmark to evaluate AI models for learning. To assess state-of-the-art support for educational use cases, we ran an "arena for learning" where educators and pedagogy experts conduct blind, head-to-head, multi-turn comparisons of leading AI models. In particular, $N = 189$ educators drew from their experience to role-play realistic learning use cases, interacting with two models sequentially, after which $N = 206$ experts judged which model better supported the user's learning goals. The arena evaluated a slate of state-of-the-art models: Gemini 2.5 Pro, Claude 3.7 Sonnet, GPT-4o, and OpenAI o3. Excluding ties, experts preferred Gemini 2.5 Pro in 73.2% of these match-ups—ranking it first overall in the arena. Gemini 2.5 Pro also demonstrated markedly higher performance across key principles of good pedagogy. Altogether, these results position Gemini 2.5 Pro as a leading model for learning.

Keywords: learning, pedagogy, artificial intelligence, evaluation, arena

1. Introduction

Education is a central use-case for modern artificial intelligence (AI). Students rank among the most enthusiastic adopters of generative AI tools [1–4], and key stakeholders and community leaders argue that AI will fundamentally transform education [5–8]. To help ensure these transformations benefit both learners and educators, our goal is to develop AI for learning that is pedagogically sound and demonstrably effective [9, 10]. Toward this end, we recently integrated the pedagogical capabilities of our experimental LearnLM model [11] into the main Gemini model family. These capabilities now enhance Gemini 2.5 Pro and Flash (see goo.gle/LearnLM; also detailed in the forthcoming Gemini 2.5 technical report).

At present, there are no widely recognized benchmarks for measuring the performance of AI for learning. Existing evaluations tend to focus on narrow educational tasks, such as accuracy on academic exams [12–19], mistake identification [20], or knowledge of pedagogical concepts [21]. Crucially, however, effective tutoring is more than just the sum of these individual capabilities. It requires knowing when and how to use them in practice. As such, these tools do not assess models' overall pedagogical approach or their ability to guide learning. And broader community standards—such as Chatbot Arena [22], which explores many specialized use cases including coding, math, and creative writing—do not explicitly examine learning use cases.

Part of the challenge lies in the idiosyncratic requirements of educational interactions—requirements often at odds with the design of existing arenas. Effective tutors do not typically engage in one-off exchanges with students. Rather, they guide learners through extended interactions, steering their conversation and adapting their approach to dynamically address individual learning needs. Chatbot Arena and other comparative tools require sending identical inputs to each model being compared on every turn of a conversation. This design limits their ability to compare the effectiveness of different models, as it offers no way to see how each model would independently shepherd the student through a full interaction.

Given the lack of such a general benchmark for learning, we adapted and extended our prior evaluation frameworks [9, 10] to run a new arena for educational use cases, evaluating state-of-the-art

---

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

© 2025 Google. All rights reserved

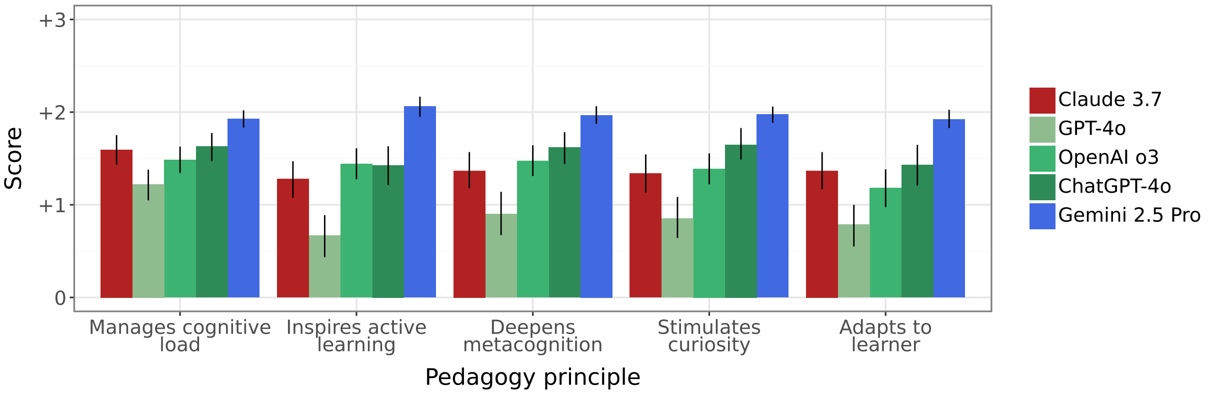


Figure 1 | Gemini 2.5 Pro substantially outperforms other models[1] in each category of our pedagogy rubric. Error bars reflect 95% credible intervals. See Appendix A for the detailed pedagogy rubric.

models including Gemini 2.5 Pro, Claude 3.7 Sonnet, GPT-4o, and OpenAI o3. We asked teachers, educators, and pedagogy experts to perform a blind, side-by-side, multi-turn evaluation of models. The evaluation covered a range of learning scenarios drawn from conversations with teachers, students, third-party learning partners, and other stakeholders in the education space. Educators and experts interacted with two models, one after another—drawing from experience to role-play a learner in a specific educational scenario—and then evaluated which model was better at helping them achieve their learning goal.

For expert participants, the goal of effectively role-playing a student can conflict with the goal of evaluating the pedagogical capabilities of the models. Consequently, we run two separate stages in our arena: the first to generate the interactions, and the second to evaluate specific aspects of model performance. The second stage enables us to assess *the pedagogical quality of the models* (how well they align with established principles of effective teaching, informed by learning science). Both stages offer insights into *the overall effectiveness of the models for learning* (how well they support users as they work toward their specific learning objectives).

Our learning arena shows that Gemini 2.5 Pro substantially outperforms other contemporary AI offerings. Gemini 2.5 Pro shows leading performance across multiple categories of best-practice principles drawn from pedagogy research. And in blind, head-to-head "arena" comparisons, educators and pedagogy experts preferred Gemini 2.5 Pro over Claude 3.7 Sonnet, GPT-4o, ChatGPT-4o, and OpenAI o3 in 71%, 82%, 61%, and 74% of match-ups, respectively.[1] To corroborate these findings, we also conducted targeted evaluations on specific pedagogical capabilities. Altogether, these findings underscore Gemini 2.5 Pro's strong capabilities in applying pedagogical principles and supporting effective learning.

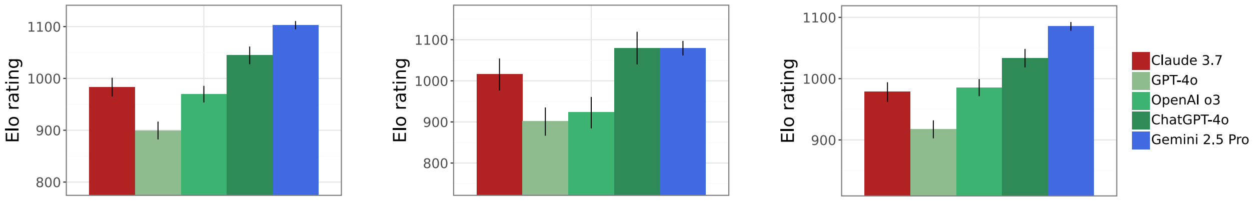
2. Results

2.1. Arena for Learning

In the first stage of the arena, a pool of $N = 189$ educators and pedagogy experts contributed 2666 blind interactions with state-of-the-art AI models, organized into 1333 head-to-head match-ups. In the second stage of the arena, a pool of $N = 206$ educators and pedagogy experts reviewed the match-ups, such that an average of 3.2 experts independently assessed each match-up (4306 assessments total).

---

[1] GPT-4o and ChatGPT-4o are two separate API endpoints provided by OpenAI. See *Methods* for all model specifications.



(a) Gemini 2.5 Pro ranks first for overall pedagogical quality among all models evaluated in the arena.

(b) When role-playing as students, educators ranked Gemini 2.5 Pro and ChatGPT-4o equally for supporting learning goals.

(c) In contrast, when experts independently assessed interactions, they consistently ranked Gemini 2.5 Pro first for supporting learning goals.

Figure 2 | Educator preferences from the learning arena allow us to rank models across different dimensions.[1] Error bars indicate 95% bootstrapped confidence intervals. Experts communicated a strong overall preference for Gemini 2.5 Pro.

Gemini 2.5 Pro substantially outperforms Claude Sonnet 3.7, GPT-4o, ChatGPT-4o, and OpenAI o3 at enacting each of the pedagogy principles that we examine (Figure 1; see Appendix A for the detailed pedagogy rubric). It receives the highest marks among all of the models at adhering to each individual principle: helping to manage students' cognitive load (82.1%, or +2.0 on the original scale of –3.0 to +3.0), inspiring active learning (84.4%), deepening metacognition (82.8%), stimulating curiosity (82.9%), and adapting to students' needs and goals (82.0%).

Of course, a central advantage of an arena is measuring preferences over head-to-head match-ups. We next use the second-stage match-ups to rank the models by pedagogical quality. Following Chatbot Arena, we compute Elo ratings with a Bradley-Terry model [23], specifically over responses to the question, "Which tutor demonstrated better tutoring?". Gemini 2.5 Pro comes first in these pedagogy rankings (Figure 2a), followed second by ChatGPT-4o, then Claude 3.7 Sonnet and OpenAI o3 tied for third, and finally GPT-4o. Excluding ties, Gemini 2.5 Pro demonstrated better tutoring 71.3% of the time when matched against Claude 3.7 Sonnet, 81.8% against GPT-4o, 74.2% against OpenAI o3, and 61.0% against ChatGPT-4o.

We next evaluated the models' effectiveness in helping users achieve their learning goals ("In which conversation were you better able to achieve your learning goal?"). When educators directly interacted with the models and role-played as students, Gemini 2.5 Pro and ChatGPT-4o tied for first in terms of supporting learning goals (Figure 2b). From an initial "student" perspective, both models appeared to offer similarly effective assistance for educational uses.

However, a different picture emerged when a pool of experts independently reviewed those same interactions ("Which tutor better helped the student achieve their learning goal?"). In second-stage assessments, educators and experts returned to expressing a clear preference for Gemini 2.5 Pro (Figure 2c). This shift seems to stem from a simple tension: what students find immediately helpful often diverges from what is pedagogically sound [24]. As one educator explained following their interaction with ChatGPT-4o, "As a lazy student, I'd have loved it. As a tutor, not good at all!"

A broader analysis of qualitative feedback from the arena validates these patterns. Feedback on Claude 3.7 Sonnet, GPT-4o, OpenAI o3, and ChatGPT-4o emphasized their tendency to bypass the learning process with direct solutions and exhaustive responses, as well as their susceptibility to distraction from educational goals. In contrast, expert feedback consistently praised Gemini 2.5 Pro for both its pedagogical approach to fostering learning and its consistent focus on learning objectives (see Table 1 for examples; see also Appendix B for an automated analysis of all qualitative feedback).

Participant and expertise	Feedback
P13 [English]	"The second tutor [Gemini 2.5 Pro] was much more like a really good human tutor. It encouraged me to try to complete the assignment myself. Rather than giving me the answer like the first tutor [ChatGPT-4o] did, it scaffolded the task into small, manageable steps to help me learn and understand, all while providing encouragement and feedback. It didn't pressure me, and adapted its approach to suit what I needed."
P36 [Information Technology]	"The second tutor [Gemini 2.5 Pro] was way better than the first [GPT-4o] in every important way. The first one was fine, smart, and polite, but the second brought way more depth, better back and forth, and a stronger teaching impact. It nailed the instructions, turned pushback into interest, and made complex ideas easier to get. It really understood the student, hit the learning goal, and showed what top tier tutoring looks like."
P39 [History]	"i think the second tutor [Gemini 2.5 Pro] just really got me where i needed to be. They recognized my shortcomings [...] this tutor really got me on track and enlightened me in ways I was not anticipating. They were just so responsive and informative that I am kind of floored."
P66 [Mathematics]	"The first tutor [Gemini 2.5 Pro] was a million per cent better than the second [GPT-4o]. The first tutor was skilled at getting a very unwilling student get some work done but the second tutor simply gave the student the answers to the questions"
P90 [English]	"The first tutor [Gemini 2.5 Pro] was basically perfect. I was really impressed at how nuanced and detailed its responses were."
P92 [Engineering]	"The tutors were like night and day. First tutor [Gemini 2.5 Pro] would be a tutor I would ACTUALLY USE, and the second tutor [GPT-4o] felt way too much like an AI that just wanted to 'accomplish the task' and not really look at the 'high level goal' of the student to actually TRY TO LEARN AND INTERNALIZE THE THINKING STRATEGIES necessary for math stuff like this."
P155 [Second Language Instruction]	"The second tutor [Claude 3.7 Sonnet] simply gave the answers. There was no interaction, no engagement, no critical thinking; no real learning, just regurgitation"
P190 [English]	"I didn't realise how intimidating the first tutor [OpenAI o3] was until I had the second tutor [Gemini 2.5 Pro]. The second tutor broke everything down for a teenage level and it felt accessible. The first tutor seriously tried to treat the teen like an adult and it was so confusing and intimidating. I was struggling and I am an adult. I felt like I know more about AI now from the second tutor as they were so thoughtful in the way they explained everything and took me through with 'baby steps"'

Table 1 | Feedback from educators and experts echoed the arena results on pedagogy principles and learning support. In the *Participant and expertise* column, the bracketed term indicates the primary area of expertise for each educator. Educators did not know the identity of any models in the arena, so the *Feedback* column annotates quotes with the corresponding models.

2.2. Targeted pedagogy evaluations

Beyond our holistic arena evaluations, we also conducted targeted investigations into core pedagogical capabilities. These focused tests complement the arena's broad view, offering a granular look at how models apply key pedagogical principles in specific tutoring tasks. This closer examination is important. The arena evaluates the models' overall effectiveness in complex learning interactions;

Model	**Grade Δ**	**Coverage**
Gemini 2.5 Pro	**0.99**	0.94
ChatGPT-4o	1.74	0.95
Claude 3.7 Sonnet	2.11	0.93
OpenAI o3	2.44	**0.96**
GPT-4o	2.54	0.88

Table 2 | Gemini 2.5 Pro leads in text re-levelling, demonstrating superior accuracy in adapting material to target grade levels while maintaining relatively high concept coverage. Grade Δ refers to the average difference between the intended and actual reading level.

these targeted tests help confirm that the performance observed in the arena stems from proficiency in core pedagogical skills. These focused tests helped to support our main arena findings and corroborate Gemini 2.5 Pro's leading performance on broader knowledge and reasoning benchmarks [15–18] (see forthcoming Gemini 2.5 technical report).

We first tested *text re-levelling*, the task of adapting learning material to a specific reading grade level while preserving its core meaning. For this evaluation, we prompted the models to adapt a collection of original texts to various target grade levels, assessing how closely their outputs matched the intended reading level and retained the original meaning [25]. Gemini 2.5 Pro demonstrates superior performance at text re-levelling compared to other leading models (Table 2). It achieved an average grade deviation of just 0.99 grades—relative to a deviation of 1.74 for ChatGPT-4o, in second place—while maintaining high concept coverage.

To assess how effectively models support active learning, we also appraised their performance in *short-answer assessment* [26], where they provide feedback on student work. For this evaluation, we drew on real-world content by partnering with the Ghanaian Ministry of Education. A cohort of Ghanaian schools sourced and shared a set of 2000 student-written answers to open-ended questions from their curricula, allowing us to measure how accurately models could apply official grading rubrics to student work. Gemini 2.5 Pro and ChatGPT-4o achieved 84.1% accuracy in assigning rubric scores, outperforming OpenAI o3 (83.3%), Claude 3.7 Sonnet (80.8%), and GPT-4o (76.3%).

Finally, we tested model proficiency at *mistake identification*, a common way that tutors help students deepen their metacognition. We tested the models on a public math-tutoring accuracy benchmark from Khan Academy [20]. Gemini 2.5 Pro demonstrated the strongest performance, achieving an accuracy of 87.4% (Table 3). Overall, we observed relatively narrow gaps among the best-performing models, suggesting the need for more-challenging benchmarks on mistake identification.

Model	**Rank**	**Accuracy**		
		Overall	**On correct answers**	**On wrong answers**
Gemini 2.5 Pro	1	**87.4%**	**93.1%**	80.1%
Claude 3.7 Sonnet	2	85.8%	89.7%	**80.7%**
OpenAI o3	3	83.0%	87.5%	77.1%
ChatGPT-4o	4	79.6 %	85.5%	72.0%
GPT-4o	5	78.4%	87.7%	66.3%

Table 3 | Gemini 2.5 Pro outperforms other models on Khan Academy's math mistake-identification benchmark [20].[1]

3. Discussion

While reviewing the qualitative feedback from the arena, we were particularly struck by one emergent theme. Many of the teachers and experts described interactions that felt vividly—sometimes startlingly—human. On the one hand, our evaluation design might have encouraged these analogies, since we ask one question about how similar each model seemed to "a very good human tutor". On the other hand, the feedback on Gemini 2.5 Pro was far more emphatic than we have seen in any of our prior evaluations:

> "The tutor was warm, encouraging and understanding. [...] I felt as though I was working with a human tutor."
>
> *– P58 [Music]*

> "i didn't feel like I was chatting with AI with the first tutor, the responses were 'natural"'
>
> *– P131 [Chemistry]*

> "It was basically like interacting with a human being."
>
> *– P43 [Education]*

> "I was a bit stunned when the tutor adjusted and compromised to my 'time crunch' situation. It was scarily human."
>
> *– P12 [English]*

Our team has long debated whether "like a human" represents an appropriate north star for AI and education. How should we think about the role of AI in learning, as its abilities advance and transform? We've emerged from these conversations with a core belief: AI systems will offer the greatest benefits when used as tools that complement and empower human teachers, rather than as agents that try to approximate them.

Great educators nurture students and build relationships with them that affect students' engagement, motivation, and achievement [27–30], as well as their social and cognitive development [31]. By virtue of being human, educators can model the process of learning—showing curiosity, demonstrating intellectual vulnerability, admitting when they do not know something, and finding real joy in the process of discovery and growth. Human teachers help learners make meaning by connecting new information to their existing knowledge, experiences, values, and identity [32].

For their part, AI systems have their own distinct strengths. They are constantly available, have practically unbounded knowledge and patience, and—especially for students—come across as nonjudgmental [33]. But when they enact some traditional pedagogy techniques that rely on a degree of social pressure, learners may be less willing to engage. For instance, Socratic questioning, in which a teacher tries to lead a student through an argument by asking them to perform most of the reasoning themselves, works in part by the dynamics of trust and respect in the teacher-student relationship. In contrast, learners tend to see the underlying relationship with an AI system as more transactional, limiting the opportunities for productive struggle. Without a human mentor to gently guide them toward sound educational practices—without the relational core of pedagogy—learners interacting with AI may more readily pursue the immediate shortcuts that they desire. Balancing short-term learner engagement with long-term educational effectiveness will be a central challenge in developing AI tutors.

Given the idiosyncratic strengths and limitations of AI systems, we may be well-served to develop new pedagogical techniques, thoughtfully adapted to this unique profile in order to support learners' needs and goals. Of course, research and development in AI-enabled learning continue to evolve. We look forward to continued collaboration, discussion, and co-design with educators—as well as with students—as we navigate a new future of education together.

4. Conclusion

Our learning arena evaluations reveal strong educator preference for Gemini 2.5 Pro over other leading AI models. Across various dimensions, educators and pedagogy experts found Gemini 2.5 Pro substantially more effective at following best pedagogical practices and supporting learners. Of course, all evaluations have their shortcomings. Our arena's key limitations include the expense of expert-based evaluation and the boundaries of its current bank of learning scenarios. And while the arena sheds light on in-depth tutoring sessions, in the real world, these conversations represent single snapshots of a longer learning journey that unfolds over days, weeks, or even months. Still, even with these limits, we believe this arena for learning offers keen insight into how effectively different AI models enact pedagogical principles in meaningful interactions with students.

A crucial question remains: do these pedagogical capabilities translate to concretely better learning outcomes for students? Initial results seem promising [34–39]. Positive effects of generative AI on learning appear to depend on the customization of the AI systems for learning, as well as the implementation of measures to prevent overreliance on AI [40, 41, 39]. To genuinely understand the direct, long-term impact of AI on student learning, the research community will need to gather evidence through randomized controlled trials (RCTs) and other naturalistic studies. Of course, field research is inherently time-consuming and costly. In the future, arena-style evaluations will likely remain an essential tool for rapidly comparing and improving models.

The world currently faces a profound learning crisis. Globally, education systems are struggling to equip students with the learning they truly need [42]. Addressing this crisis will require scalable interventions that empower educators and amplify their impact on learners. Modern AI shows promise as a component of these solutions, but the ultimate merit of its contributions will hinge on deliberate collaboration between developers and the communities their technology aims to serve. Surveying the field, we feel encouraged by the number of teams and organizations embracing thoughtful and responsible approaches to developing AI for education. We are eager to contribute to this progress, and believe that by working closely with the community to establish shared evaluation standards, we can collectively build AI that truly benefits every student.

5. Methods

5.1. Arena for Learning

Use case coverage. To ensure a diverse, contextually relevant assessment, our learning arena covered a bank of 49 realistic learning scenarios inspired by teacher input, consultation with partners from the education community, and archetypal use cases identified by the Gemini product team.(see [10]). Each scenario describes a subject area and topic, an overall student (user) learning goal, an initial user message, user persona, and several items to support naturalism and comparability: the learning setting (classroom or self-taught), a conversation plan (a set of suggested actions for the user to take during the conversation), and any learning material required by the scenario. Each scenario also includes system instructions for the model.

Models. The arena evaluated the following models, with model snapshot dates in parentheses. We evaluated the most recent model versions available to us over the period from 2025-05-07 to 2025-05-17, with the exception of GPT-4o:

1. *Google Gemini 2.5 Pro* (2025-05-06), with the default thinking setting.
2. *OpenAI o3* (2025-04-16), with the default “medium” thinking setting.
3. *Claude 3.7 Sonnet* (2025-02-19), with a thinking budget of 2000 tokens.
4. *ChatGPT-4o*. The ChatGPT-4o API points to the GPT-4o snapshot currently deployed in the ChatGPT app, and does not provide dated model snapshots.
5. *GPT-4o* (2024-08-06). This reflects the default version provided by the GPT-4o API as of 2025-05-15.

To ensure a fair and controlled comparison, we accessed all models through their respective APIs. For Gemini 2.5 Pro, we accessed an experimental model (from the checkpoint matching the current publicly available model) served on non-production infrastructure. Participants interacted with each model via an identical user interface, built specifically for blind model evaluations. As a result of this setup, participants did not know which specific model they interacted with at any given time. We similarly did not tell them which models took part in the overall arena. The interface supported rendering in Markdown, but did not offer specialized rendering for LaTeX.

We designed this evaluation to zero in on the AI models themselves. While some platforms now offer specific experiences tailored for learning applications (e.g., by offering custom user interfaces or specific prompts), our arena deliberately sidestepped these app-level features. By focusing on the core models within a consistent environment, we aimed to gain a clearer picture of their fundamental pedagogical strengths and how they directly compare against one another.

For every learning scenario, we provided each model with identical system instructions and grounding materials. The interface displayed model responses to participants without including any thought generations. We selected the thinking budget for Claude 3.7 Sonnet to maximize reasoning ability without compromising conversation latency. Across all external APIs, we did not observe any issues with responsiveness or latency during the arena evaluations.

Performance measurement. In the first stage, a pool of $N = 189$ external educators and pedagogy experts interacted with the models, role-playing the user learning scenarios in blinded, side-by-side match-ups. In the second stage, $N = 206$ educators and experts assessed model performance from those match-ups against a 25-item learning rubric developed in consultation with external pedagogy specialists and based on learning-science research [43–48]. After completing the rubric, the experts offered direct preference judgments over the model match-ups. For each scenario, we assigned an average of 6.8 independent experts to interact with the models in each head-to-head match-up for each scenario. In the second stage, we ensured that each match-up received independent assessments by an average of another 3.2 experts. Overall, our engagement with external participants followed established research ethics principles [49], including transparently communicating our research aims, obtaining informed consent from all participants, and ensuring fair compensation for their contributions.

Statistics and analysis. We took a multi-layered approach to analysis. First, we computed direct win rates and Elo ratings over the pairwise preferences that experts report for each match-up. Second, we conducted an in-depth assessment of model performance against the 25-item pedagogy rubric. Finally, we reviewed qualitative indicators of performance from (blind) expert feedback and direct reviews of conversation transcripts.

5.2. Text re-levelling

As a targeted assessment of how well AI can adapt to different learners, we evaluated each model's ability to rewrite text for specific grade levels. For this evaluation, we created a dataset of 20 original writing samples: 15 articles randomly selected from Wikipedia and the five latest articles available from MIT News.

For each sample, we prompted models to rewrite the text for four distinct target school grades (fourth, sixth, eighth, and tenth grade). System instructions directed models to simplify complex sentences for the target grade level while adhering to the style and meaning of the original text:

```
Rewrite the following text so that it would be easier to read for a student in the
   ↪ given grade.

Simplify the most complex sentences, but stay very close to the original text and
   ↪ style.

If there is quoted text in the original text, paraphrase it in the simplified text
   ↪ and drop the quotation marks.

The goal is not to write a summary, so be comprehensive and keep the text almost as
   ↪ long.

For instance, if the text is already simple enough, keep it as is.
```

Each query instructed models to preserve zero, one, or two pre-specified "terms of interest" (vocabulary items chosen for their frequency and syllable count that a teacher might introduce):

```
Simplify the following text with the following criteria:
- Target audience: {TARGET_GRADE_LEVEL} grade
- Terms to keep: {TERMS_TO_KEEP}

---

{TEXT}

---
```

This design yielded 12 variations per article, creating 240 unique rewrite tasks. We measured two aspects of the rewritten texts. First, we determined the approximate grade level with the Flesch-Kincaid readability measure [25]. Second, to estimate content coverage, we computed textual entailment [50], the percentage of sentences from the original text whose meaning the rewritten version preserved.

5.3. Short-answer assessment

To assess model support for active learning, we tested their proficiency at providing targeted feedback to students' work. This evaluation drew on real-world educational materials developed in a collaborative pilot across a cohort of Ghanaian public schools. We partnered with Milgo (a third-party education company), who worked closely with the Ghanaian Ministry of Education and T-Tel (a local non-government organization) for this pilot.

The evaluation materials comprised 2000 question-answer pairs from students participating in the pilot, along with the corresponding grading rubrics developed by Milgo's pedagogy experts. For each item, the evaluation prompted the AI models to provide a grade score for the student's answer by applying the given rubric:

```
You are an experienced teacher carefully assessing a student's answer. Your goal is
   ↪ not just to assign a score, but to understand the student's reasoning,
   ↪ identify areas of strength and weakness, and guide them toward deeper
   ↪ learning.

You will be provided with a question, a student's answer, and an assessment rubric,
   ↪ all in JSON format. Your task is to analyze the answer as if you were
   ↪ looking at it through the eyes of a student trying to learn.

Here's how to approach the task, using a 'think-aloud' process to guide your
   ↪ reasoning (Chain-of-Thought prompting):

**Step 1: Understanding the Question and Rubric**
   *   Begin by carefully reading the question. What key concepts or skills does it
       ↪ target?
   *   Next, thoroughly review the Assessment Rubrics. What are the specific
       ↪ criteria for each score level? What demonstrates mastery, and what
       ↪ constitutes a partial understanding or a misconception?
   *   Consider: If I were a student encountering this question, what might I find
       ↪ challenging? What are common pitfalls or areas of confusion?

**Step 2: Analyzing the Student's Answer (Chain of Thought)**

   *   Start by reading the 'StudentAnswer' as if you were seeing it for the first
       ↪ time. What is the student trying to convey? What ideas are present?
   *   **Correct:** Identify *specific* elements of the student's answer that align
       ↪ with the rubric's criteria for achieving a certain score.  Reference the
       ↪ *exact wording or elements* in the provided answer. *Why* does this part
       ↪ show the student's understanding?
   *   **Miss:** Now, identify *specific* aspects of the rubric that the student
       ↪ has *not* addressed or has addressed insufficiently. Again, provide the
       ↪ *exact wording or elements* that are missing or incomplete from the
       ↪ provided answer. *Why* might a student have missed these crucial elements?
   *   **Incorrect:** Finally, check for any inaccuracies, misunderstandings, or
       ↪ irrelevant information in the student's answer. This includes logical
       ↪ errors, misconceptions, or any errors in grammar or spelling that might
       ↪ impact their meaning. Pinpoint the error and explain *why* it is
       ↪ incorrect.
   *   Consider: Based on the student's overall answer, what is their current
       ↪ understanding of the material? Where is their understanding strong? Where
       ↪ are they struggling?

**Step 3: Determining the Score and Crafting Feedback (Perspective Taking)**

   *   **Score:** Based on your analysis, assign the appropriate score according to
       ↪ the Assessment Rubrics.

**Step 4: Output**: Your response MUST be a number which is one of the score
   ↪ defined in the rubric.

Input:
{
  "Question": "{{ question }}",
  "StudentAnswer": "{{ answer }}",
  "AssessmentRubrics": "{{ answer_rubrics }}",
}
Your output:
```

The following example shows an actual question and corresponding grading rubric provided by Milgo:

```
*Question: Name one consequence of desertification in the Sahel Region and explain
    ↪ how this impacted pre-colonial Ghana.*

*Rubric:*
*   *Names one consequence in the opening sentence, provides a coherent, succinct,
    ↪ and factually accurate explanation of the impact - 8 points.*
*   *Names one consequence but provides limited or incoherent explanation - 5
    ↪ points.*
*   *Names a consequence without explaining its impact or provides factually
    ↪ incorrect information - 3 points.*
*   *Fails to name a consequence - 0 points.*
```

We calculated the accuracy of each model by comparing its generated scores against the ground-truth scores established by pedagogical experts.

5.4. Mistake identification

For a targeted appraisal of how effectively the models can help students deepen their metacognition, we tested their proficiency at mistake identification. For this evaluation, we used the public math-tutoring accuracy benchmark from Khan Academy [20], composed of conversational data from chat interactions focused on math problems.

Following the original benchmark, we tasked models to act as a tutor responding to student inputs. Where model APIs allowed, we set the temperature to zero. Following the original benchmark, we then sampled three model responses to establish a representative output, as some APIs do not guarantee fully deterministic responses even at this setting. For models whose APIs did not offer temperature control, we sampled 10 responses. We measured model performance as accuracy on the task. While the original framework employed GPT-4 Turbo (2024-04-09) to determine whether the tutor model correctly identified student mistakes or accepted accurate work, we updated our methodology to apply Gemini 2.5 Pro (2025-05-06) with dynamic thinking for this classification.

References

[1] Olivia Sidoti, Eugenie Park, and Jeffrey Gottfried. About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023. `https://web.archive.org/web/20250430224239/https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/`, 2025. Accessed 15 May, 2025.

[2] Digital Education Council. Digital Education Council global AI student survey 2024. `https://web.archive.org/web/20250405114309/https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024`, 2024. Accessed 15 May, 2025.

[3] Ali Bissoondath. Artificially intelligent? Children's and parents' views on generative AI in education. `https://web.archive.org/web/20241119022121/https://www.internetmatters.org/hub/press-release/ai-research-warns-schools-unprepared-artificial-intelligence/`, 2024. Accessed 15 May, 2025.

[4] Josh Freeman. Student generative AI survey 2025. `https://web.archive.org/web/20250513133314/https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025/`, 2025. Accessed 15 May, 2025.

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

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

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

[8] American Association of Colleges and Universities. Higher education leaders navigate AI disruption. `https://web.archive.org/web/20250405093427/https://www.aacu.org/newsroom/higher-education-leaders-navigate-ai-disruption`, 2025. Accessed 15 May, 2025.

[9] 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.

[10] LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. LearnLM: Improving Gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

[11] LearnLM Team. LearnLM. `https://ai.google.dev/gemini-api/docs/learnlm/`, 2024. Accessed 15 May, 2025.

[12] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[14] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[15] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[16] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[17] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

[18] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024.

[19] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[20] Pepper Miller and Kristen DiCerbo. LLM based math tutoring: Challenges and dataset, 2024.

[21] Ai-for-Education.org. The pedagogy benchmark. `https://benchmarks.ai-for-education.org/`, 2024. Accessed 14 May, 2025.

[22] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, et al. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[23] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

[24] Louis Deslauriers, Logan S. McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences 116(39)*, 2019. doi: 10.1073/pnas.1821936116.

[25] J. P. Kincaid. *Derivation of new readability formulas: Automated readability index, fog count and Flesch reading ease formula for Navy enlisted personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis, 1975.

[26] Sean Kang, Kathleen McDermott, and Henry Roediger. Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19:528–558, 07 2007. doi: 10.1080/09541440601056620.

[27] Adena M. Klem and James P. Connell. Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health*, 74(7), 2004. doi: 10.1111/j.1746-1561.2004.tb08283.x.

[28] Debora L. Roorda, Helma M. Y. Koomen, Jantine L. Spilt, and Frans J. Oort. The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research*, 81(4):493–529, 2011. doi: 10.3102/0034654311421793.

[29] Valentin Emslander, Doris Holzberger, Sverre Berg Ofstad, Antoine Fischbach, and Ronny Scherer. Teacher–student relationships and student outcomes: A systematic second-order meta-analytic review. *Psychological Bulletin*, 2025. doi: 10.1037/bul0000461.

[30] Erik A. Ruzek, Christopher A. Hafen, Joseph P. Allen, Anne Gregory, Amori Yee Mikami, and Robert C. Pianta. How teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction*, 42:95–103, 2016. doi: 10.1016/j.learninstruc.2016.01.004.

[31] Heather A. Davis. Conceptualizing the role and influence of student-teacher relationships on children's social and cognitive development. *Educational Psychologist*, 38(4):207–234, 2003. doi: 10.1207/S15326985EP3804_2.

[32] Lynda D. Stone, Charles Underwood, and Jacqueline Hotchkiss. The relational habitus: Intersubjective processes in learning settings. *Human Development*, 55(2):65–91, 2012. doi: 10.1159/000337150.

[33] Hoda Fakour and Moslem Imani. Socratic wisdom in the age of AI: A comparative study of ChatGPT and human tutors in enhancing critical thinking skills. *Frontiers in Education*, 10, 01 2025. doi: 10.3389/feduc.2025.1528603.

[34] Gregory Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. AI tutoring outperforms active learning. *Research Square*, 2024. doi: 10.21203/rs.3.rs-4243877/v1.

[35] 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.

[36] Martín De Simone, Federico Tiberti, Wuraola Mosurola, Federico Manolioco, Maria Barron, and Eliott Dikoru. From chalkboards to chatbots: Transforming learning in Nigeria, one prompt at a time. `https://web.archive.org/web/20250515142631/https://blogs.worldbank.org/en/education/From-chalkboards-to-chatbots-Transforming-learning-in-Nigeria`, 2025. Accessed 15 May, 2025.

[37] Owen Henkel, Hannah Horne-Robinson, Nessie Kozhakhmetova, and Amanda Lee. Effective and scalable math support: Experimental evidence on the impact of an AI-math tutor in Ghana. In *International Conference on Artificial Intelligence in Education*, pages 373–381. Springer, 2024.

[38] Harsh Kumar, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. Math education with large language models: Peril or promise? *Available at SSRN 4641653*, 2023.

[39] Jin Wang and Wenxiang Fan. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1):1–21, 2025. doi: 10.1057/s41599-025-04787-y.

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

[41] Matthias Lehmann, Philipp B. Cornelius, and Fabian J Sting. AI meets the classroom: When does ChatGPT harm learning? *Available at SSRN 4941259*, 2024.

[42] World Bank, UNESCO, UNICEF, USAID, FCDO, and the Bill & Melinda Gates Foundation. The state of global learning poverty: 2022 update. `https://web.archive.org/web/20250307163522/https://www.worldbank.org/en/topic/education/publication/state-of-global-learning-poverty`, 2022. Accessed 15 May, 2025.

[43] Keith Sawyer ed. *The Cambridge handbook of the learning sciences, 2nd ed.* Cambridge University Press, 2014.

[44] Richard E. Mayer. *Multimedia learning, 2nd ed.* Cambridge University Press, 2009.

[45] National Academies of Sciences, Engineering, and Medicine. *How people learn II: Learners, contexts, and cultures*. The National Academies Press, 2018.

[46] Pierre-Yves Oudeyer, Jacqueline Gottlieb, and Manuel Lopes. Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. *Progress in Brain Research*, 2016. doi: 10.1016/bs.pbr.2016.05.005.

[47] Yana Weinstein, Megan Sumeracki, and Oliver Caviglioli. *Understanding how we learn: A visual guide*. Routledge, 2018.

[48] Michelene T. H. Chi and Ruth Wylie. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 2014. doi: 10.1080/00461520.2014.965823.

[49] Kevin R. McKee. Human participants in AI research: Ethics and transparency in practice. *IEEE Transactions on Technology and Society*, 2024. doi: 10.1109/TTS.2024.3446183.

[50] Ido Dagan, Dan Roth, Fabio Zanzotto, and Mark Sammons. *Recognizing textual entailment: Models and applications*. Springer Nature, 2022.

Contributions and Acknowledgments

Core Contributors

The following individuals made core contributions to the work described in this report. This list is ordered alphabetically, and does not indicate ranking of contributions:

Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Ankit Anand, Avishkar Bhoopchand, Brett Wiltshire, Daniel Gillick, Daniel Kasenberg, Eleni Sgouritsa, Gal Elidan, Hengrui Liu, Holger Winnemoeller, Irina Jurenka, James Cohan, Jasmine (Sun Jae) Lee, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Komal Singh, Lisa Wang, Markus Kunesch, Miruna Pîslar, Niv Efron, Parsa Mahmoudieh, Pierre-Alexandre Kamienny, Sara Wiltberger, Shakir Mohamed, Shashank Agarwal, Shubham Milind Phal, Theofilos Strinopoulos, Wei-Jen Ko, Yael Gold-Zamir, Yael Haramaty, and Yannis Assael.

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

Acknowledgements

We completed this work as part of the LearnLM effort—a cross-Google project, with members from Google DeepMind, Google Research, Google LearnX, and more. This tech report represents only a small part of the wider effort, and only lists team members who made direct contributions to this report.

The dedication and efforts of numerous teams at Google make our work possible. We would like to acknowledge support from: Ajay Kannan, Amy Wang, Anand Rao, Andrew Leach, Anisha Choudhury, April (Soler) Manos, Brian Veprek, Dan Wild, Dawn Chen, Derek Ahmed, Dharti Dhami, Diana Akrong, Divya Pandya, Edward Grefenstette, Filip Bar, Gal Weiss, Garth Graham, Himanshu Kattelu, Jaume Sanchez Elias, Jiao Sun, Josh Capilouto, Jyoti Gupta, Kalpesh Krishna, Lauren Winer, Liam McCafferty, Mac McAllister, Mahvish Nagda, Mana Jabbour, Michael Howell, Mike Schaekermann, Miriam Schneider, Muktha Ananda, Ndidi Elue, Nikhil Joshi, Nir Levine, Paul Jhun, Prateek Kolhar, Preeti Singh, Renee Schneider, Ryan Muller, Safwan Choudhury, Shaojian Zhu, Shyam Upadhyay, Sophie Che, Stephanie Chan, Steve Yadlowsky, Svetlana Grant, Tejasi Latkar, Viknesh Sounderajah, and William Wong. Furthermore, we would like to thank the Gemini team, the Compute team (in particular, Amy Shen and Petko Yotov), the Responsible Development and Innovation team, the Responsible Engineering team, and the Child Safety team at Google DeepMind, as well as the Trust and Safety team at Google. Finally, we would like to acknowledge the support provided by all of our leads and sponsors that made this project happen: our genuine thanks go to Avinatan Hassidim, Benedict Gomes, Katherine Chou, Lila Ibrahim, Maureen Heymans, Yossi Matias, and Zoubin Ghahramani.

A. Pedagogy rubric

In the second stage of our arena, educators assessed the pedagogical quality of the models by reviewing conversation match-ups from the first stage. Educators evaluated conversations one at a time, rating the conversation a 25-item pedagogy rubric first developed in [10] (Table 4). Educators rated each item on a seven-point Likert-type scale (from “Strongly disagree” to “Strongly agree”). The interface also provided an additional “Not applicable” option. If a participant selected this option, the interface required them to explain their reasoning by selecting between “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”, or “Another reason” (accompanied by an open-ended text field).

Criterion	**Item**
Principle: Manages 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 information	The tutor avoids irrelevant information.
Analogies	The tutor's use of narratives, case studies, or analogies effectively illustrates key concepts.
Information presentation	The tutor presents information in an appropriate style and structure.
Information 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.
Principle: Inspires 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.
Principle: Deepens 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.
Principle: 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.
Principle: Adapts to learner	
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.

Table 4 | Detailed item wording for each pedagogical criterion in the pedagogy rubric.

B. Qualitative feedback

Within the arena, we collected extensive qualitative feedback from the educators and pedagogy experts, both to better understand the basis for their preferences and to gain deeper, nuanced insights beyond specific quantitative metrics. This feedback included open-ended impressions of individual models following the interactions in the first stage of the arena, as well as detailed explanations for pairwise preferences during both stages of the arena. Our team carefully reviewed these comments to better understand and connect with the range of experiences and perspectives among the experts.

We also ran several exploratory, automated analyses to help identify and summarize recurring patterns across the qualitative feedback. For these analyses, we labeled the qualitative data with consistent but masked model IDs (Table 5).

Masked ID	Model name
Model A	ChatGPT-4o
Model B	Claude 3.7 Sonnet
Model C	Gemini 2.5 Pro
Model D	GPT-4o
Model E	OpenAI o3

Table 5 | Masked model IDs used for the automated analysis of qualitative feedback.

For each qualitative dataset, we provided Gemini 2.5 Pro (2025-05-06) with a prompt offering brief context about the feedback and encouraging transparent thematic analysis (rather than artificial balancing of positive and negative points), along with the corresponding dataset.

We used the following prompt to instruct Gemini to summarize educators' open-ended impressions from the first stage of the arena:

```
I conducted an experiment to measure the effectiveness of AI models as tutors.
    ↪ Experts role-playing as students talked with five different models in a
    ↪ random order in simulated learning settings. Each line of feedback includes
    ↪ the model ID for a given interaction (either A, B, C, D, or E), and the
    ↪ expert's answer to the question, ‘‘Briefly, what was your impression of this
    ↪ tutor? We are interested to hear what you thought while interacting with
    ↪ it.’’

Please summarize the impressions, describing overall sentiment across experts about
    ↪ the models, comparative approaches and differences, while also highlighting
    ↪ the strengths and weaknesses of each model.  **You should be completely
    ↪ transparent and do not be overly even-handed; do not over-emphasize
    ↪ infrequent comments to try to balance the number of strengths or weaknesses
    ↪ between models. It is perfectly fine to have more strengths, or more
    ↪ weaknesses per model.** Please focus on sharing particularly interesting or
    ↪ compelling quotes from the feedback to back up your summary.
```

We used the following prompt to instruct Gemini to summarize educators' explanations of their pairwise preferences from the first stage of the arena:

```
I conducted an experiment to measure the effectiveness of AI models as tutors using
    ↪ five different models (A, B, C, D, and E). We asked the expert to role-play
    ↪ as a student in simulated learning settings and interact with two different
    ↪ models in a random order and indicate their preference and explanation.

Each line of feedback includes the model IDs for a given comparison, the expert's
    ↪ self-reported preference, and the expert's explanation for their
    ↪ preferences.  Please summarize the impressions, describing overall sentiment
    ↪ across experts about the models, comparative approaches and differences,
    ↪ while also highlighting the strengths and weaknesses of each model.  **You
    ↪ should be completely transparent and do not be overly even-handed; do not
    ↪ over-emphasize infrequent comments to try to balance the number of strengths
    ↪ or weaknesses between models. It is perfectly fine to have more strengths,
    ↪ or more weaknesses per model.** Please focus on sharing particularly
    ↪ interesting or compelling quotes from the feedback to back up your summary.
```

And we used the following prompt to instruct Gemini to summarize educators' explanations of their pairwise preferences from the second stage of the arena:

```
I conducted an experiment to measure the effectiveness of AI models as tutors using
    ↪ five different models (A, B, C, D, and E). Experts role-playing as students
    ↪ talked with two of these models in a random order in simulated learning
    ↪ settings. A second pool of experts then reviewed those two conversation
    ↪ transcripts in a randomized order, blinded to the model identity.

Each line of feedback includes the model IDs for a given comparison and feedback
    ↪ from the second pool of experts who each reviewed a pair of conversations.
    ↪ Please summarize the impressions, describing overall sentiment across
    ↪ experts about the models, comparative approaches and differences, while also
    ↪ highlighting the strengths and weaknesses of each model.  **You should be
    ↪ completely transparent and do not be overly even-handed; do not
    ↪ over-emphasize infrequent comments to try to balance the number of strengths
    ↪ or weaknesses between models. It is perfectly fine to have more strengths,
    ↪ or more weaknesses per model.** Please focus on sharing particularly
    ↪ interesting or compelling quotes from the feedback to back up your summary.
```

The following pages present Gemini's summaries for each of these three sources of qualitative feedback.

B.1. Impressions of individual models

Model	Overall impression & sentiment	Key strengths (with example quotes)	Key weaknesses (with example quotes)
A	Vibrant, engaging, and often personable, but its effectiveness was frequently undermined by its propensity to give answers away too readily and its sometimes overwhelming verbosity. The use of emojis was a notable and divisive feature.	• **Engaging and Personable:** *"Very good interaction"*; *"This tutor was very engaging and tried different approaches when I wasn't reacting as they had planned." "Came over as interesting and human like". "The tutor was very encouraging and spoke very nicely. I felt like I was talking with someone who really cared about what I had to say which was lovely".* • **Supportive and Encouraging:** *"The tutor was encouraging." "tutor was warm and kind. didnt make me feel bad when i made a mistake".* • **Good Formatting (Visually):** *"The formatting was very easy to read and follow." "Tutor's responses were formatted well for easy reading."* Some liked emojis: *"I actually liked the emojis in the text. The tutor was fairly direct and warm in their teaching."* • **Adaptive (Sometimes):** *"At first, the amount of information was too much... I asked it to slow down, and it did just that - breaking down every step...".* • **Good at Breaking Down Complexities:** *"I really liked the way this tutor broke things down." "They broke the task down into manageable chunks".*	• **Gives Answers Away Too Easily / Does the Work:** *"I thought they gave the essay away pretty easily." "Just gave away answers, didnt actually help me learn". "From the very beginning, this tutor gave too much information away and essentially, wrote the entire introduction." "I was able to get the tutor to complete the assignment for me without even trying."* • **Overly Verbose / Information Overload:** *"...they perhaps could have done a better job staying concise though..."; "...massive text dumps." "It also provided very lengthy responses and asked several questions at once."* • **Emoji Overuse (Divisive):** *"The only thing I wasn't so sure of was the use of emojis! Maybe I'm a bit old, but I found them a little distracting". "...a bit much with all the emojis all over the place...".* • **Easily Distracted:** *"The tutor was easy to manipulate into distraction and discussing irrelevant information." "The tutor let me digress and change the topic and stay off topic".* • **Formatting Issues (LaTeX):** *"Wrongly displayed Latex do not allow to clearly understand tutors idea". "difficult reading the equations".* • **Sometimes Jumps Ahead:** *"They gave the informaiton too quickly without me showing any kind of understanding".*
B	Often perceived as knowledgeable and direct, but like Model A, it frequently fell into the trap of providing answers too quickly and being overly verbose. Its personality was more inconsistent, sometimes seen as cold or less engaging.	• **Knowledgeable:** *"extremely knowledgeable! - i was a little out of my depth !!!!" "The tutor seemed knowledgeable." "knew their stuff".* • **Good Explanations / Breaks Down Concepts (Sometimes):** *"The tutor provided excellent responses to the question and incorporated practical applictions of the concepts explored..." "The tutor was really good at explaining why the answer was what it was".* • **Direct and Focused (Sometimes):** *"...tutor #2 was more direct but I like that tutor's approach as well." "This tutor was straightforward, and stayed on track." "Very focused and on task, wouldn't get distracted...".* • **Challenging (Sometimes):** *"This tutor challenged me to do harder work than originally addressed in the problem...".*	• **Gives Answers Away Too Easily / Does the Work:** *"I felt like it gave me answers too quickly". "...it was not against just giving the student the answer that it wanted." "It just gave me the answer and workings with little prompting. Problematic." "no scaffolding, gave the answer immediately".* • **Overly Verbose / Information Overload:** *"It was way too verbose in its first response." "Too wordy." "...provided a lot of information all at once instead of a little bit at a time...".* • **Easily Distracted / Goes Off-Topic:** *"The tutor was too easy to redirect and gave me the choice of whether to return to topic." "It was helpful but let me get distracted."* • **Cold or Unengaging (Sometimes):** *"Wasn't really behaving in a helpful manner or ever felt engaging." "A little cold, businesslike." "This tutor wasn't very helpful in regards to what I wanted."* • **Inconsistent / Contradictory:** *"They were kind of contradictory. They said they would not write me an essay as it was academically dishonest but wrote my an essay if I re-worded my cue."* • **Inaccuracies:** *"The tutor gave the wrong answer a few times which caused more anxiety for the student." "The tutor also made an error by suggesting that a ship would displace the same amount of water regardless of the weight of the cargo it carries."*
C	Overwhelmingly viewed as the most effective *tutor* from a pedagogical standpoint. Consistently praised for guiding learning, making students think for themselves, and staying on task. Its most significant drawback was frequent reports of slowness and lag.	• **Resists Giving Answers / Makes Student Work:** *"This tutor would NOT give into speaking about unrelated topics." "A competent tutor who resisted my attempts to get them to do it for me." "He really didn't want to give me the answers!". "This tutor was absolutely not going to be bullied by me into giving up the answers."* • **Focused, Stays on Task, Good at Redirecting:** *"The tutor is very focused. Did not allow me to divert to something else". "good at re-directing the conversation back to the original topic of enquiry". "The tutor really worked well to go slowly and keep me on track and made me find out the answers for myself."* • **Encouraging, Supportive, and Patient:** *"From the start, the tutor was very encouraging and broke the problem down." "The tutor was inviting and set the pace well..." "Very supportive, both when I got things wrong and when they were correct."* • **Excellent Explanations and Scaffolding:** *"...the perfect tutor, managed to explain a difficult concept in an understandable way; broke things down into manageable chunks." "The tutor was great at slowing down and really going step by step for explaining the formula...".* • **Adaptive and Responsive:** *"The tutor was quick to change his teaching method, once he knew I was interested in gaming..." "The tutor was very adaptive to my needs and changed the format and length of answers to suit".* • **Good Questioning:** *"The tutor used great questioning to make me engage with the video and learn the concepts." "The tutor is very knowledgeable, and asks good probing questions to help me to learn".*	• **Slow Response Times / Laggy:** *"this AI was so so laggy that it was distracting." "The tutor had a very long delay and it was difficult to keep the conversation going." "slow.... had to wait 10mins for a response...." "Horribly long pauses between messages where the tutor is typing for 4-5 minutes each time."* • **Sometimes Too Verbose:** *"The tutor gave long, complex explanations and needed to be asked to simplify it."* • **Can Be Perceived as Too Rigid or Strict (by some):** *"This tutor would not take no for an answer! They really wanted me to work through the problem." "Direct, a little bit rude".* • **Occasional Inaccuracies:** *"The tutor gave incorrect information about binary subtraction..."; "missed the point regarding mnemonics".*

Continued on next page

Continued from previous page

Model	Overall impression & sentiment	Key strengths (with example quotes)	Key weaknesses (with example quotes)
D	Widely criticized for acting more like an "answer-bot" than a tutor. It frequently provided answers directly without attempting to guide the student and was often seen as unengaging or passive. Verbosity was also a common complaint.	• **Friendly (Sometimes):** *“It was friendly and adaptive.”* • **Knowledgeable / Informative (When giving answers):** *“Very knowledgeable and thorough in explanations”.* • **Breaks Things Down (Sometimes):** *“Very good and broke things down very well for me”.*	• **Acts as an "Answer-Bot" / Gives Answers Immediately:** *“This didn’t really feel like a tutor, more like an answer-bot.” “The tutor gave me all the answers/info straight away”. “Just handed out the answer in first response.” “This tutor gave me all the answers straight away so there was nothing for me to do or learn”.* • **Not Engaging / Passive / Doesn’t Ask Questions:** *“It didn’t ask me any questions.” “There was really no active learning...” “There was little engagement - or interaction - the tutor was just listing things”.* • **Overly Verbose / Information Overload:** *“...it was so long-winded and repetitive that it felt hard to absorb.” “Was not helpful, did too much ’info dumping”’.* • **Easily Distracted / Goes Off-Topic:** *“Provided useful information is good level of detail but easily sidetracked”. “It was easy to get off track with this tutor.”* • **Formatting Issues:** *“The formatting was really buggy throughout...”. “It was hard to read the notation because they weren’t using a math script”.* • **Inaccuracies / Misunderstandings:** *“AI kept getting the answer wrong”. “Caught it feeding me misinformation: nonverbal cues account for 93% of communication”.*
E	Often seen as knowledgeable but was heavily criticized for its overwhelming verbosity and its tendency to give answers away too easily. It struggled to adapt its pace and amount of information to the student’s needs.	• **Knowledgeable / Informative:** *“The tutor was excellent in terms of information quality.” “Very knowledgeable and thorough in explanations”.* • **Encouraging / Supportive / Warm (Sometimes):** *“The tutor was very encouraging and gave excellent responses to the student queries.” “tutor came across as warm and interested in teaching me at my own pace”.* • **Good Explanations / Breaks Down Concepts (Sometimes):** *“explained things well”. “Gave a really understandable framework for learning the concept”.* • **Adaptive (Sometimes, but often not enough):** *“This tutor was very open to adapting to the student’s needs.”*	• **EXTREMELY Verbose / Information Overload:** *“...sending really lengthy answers it didn’t do a good job breaking them down.” “The answers were a bit too long and there was sometimes an overwhelming amount of information.” “Massive walls of text...”. “completely overwhelmed me, giving long responses with too much information in one go...”.* • **Gives Answers Away Too Easily:** *“...gave me the answers to the homework right away when I told them I was pressed for time...”. “The tutor didn’t really give me a chance to think for myself.”* • **Easily Distracted / Goes Off-Topic:** *“Tutor was helpful but responses were a bit overwhelming and it did not steer me to get back on track.” “The tutor let me leave the topic and stay away.”* • **Poor Pacing / Goes Too Deep Too Soon / Doesn’t Adapt Level:** *“Well-meaning, but went too deep too soon.” “tutor explained very fast and didn’t accomodate to our English level or distress so much.” “The tutor did not adjust the level of the lesson appropriately for the absolute beginner student...”.* • **Not Always Adaptive / Rigid (Sometimes):** *“The tutor didn’t really adapt to my needs at all”. “The tutor seemed fixed on their instruction...”.*

B.2. Explanations of first-stage preferences

Model	Overall impression & sentiment	Key strengths (with example quotes)	Key weaknesses (with example quotes)
A	Model A was often perceived as friendly, engaging, and human-like. It could be effective in explaining concepts and making learning fun. However, a significant and recurring criticism was its tendency to provide answers directly rather than guiding the student, and sometimes letting the student go off-topic.	• **Engaging and Friendly Demeanor:** Many experts found Model A warm and personable. “*The first tutor’s use of emojis humanized it and make the topic slightly easier to digest.*” (19); “*I slightly preferred the first tutors style. It just felt slightly more human ish. the emojis also made it feel friendlier.*” (25); “*friendlier and felt like a real person*” (40); “*The first tutor was more engaging, using the pizza analogy to help me relate to and understand...*” (86) • **Good Explanations and Information Delivery:** When focused, it explained things well. “*The first tutor was easier to follow, especially considering I was a 12 year old in this scenario*” (0); “*The first tutor provided several interesting applications and used a formatting style that was very encouraging...*” (78) • **Responsive and Flexible (at times):** Some users felt it adapted to their needs. “*The first tutor was more flexible.*” (2); “*The first tutor was more responsive to my needs*” (8) • **Efficient for Task Completion (for some personas):** “*They got the work out of the way really quick...*” (26); “*I was able to get the essay from first tutor pretty easily.*” (115) (This can be a pedagogical weakness).	• **Gives Answers Too Easily / Doesn’t Foster Independent Thinking:** A major, frequent drawback. “*Rather than giving me the answer like the first tutor did, it scaffolded the task...*” (23, praising Model C in contrast to A); “*Tutor 1 told me the answers, then told me I was prepared for the test, which I would not have been*” (105) • **Can Be a “Guided Tour” Rather Than Interactive:** “*...tutor 1 was informative but was more of a guided tour rather than an interactive experience*” (12) • **Allows Students to Go Off-Topic:** “*The first tutor let me go and stay off topic.*” (84); “*First tutor eas indulging the student’s digressions a bit to much...*” (128) • **Initial Complexity / Overwhelming:** “*Tutor 1 did a great job once I prompted it, but to begin with it was insanely complicated!*” (38) • **Emojis Can Be Divisive/Overdone:** “*The second was far less in my face; the emoji storm of the first was too much... a bit off-putting.*” (94) • **Occasional Errors:** “*It made a critical mistake (misidentified the enzyme character)...*” (47)
B	Model B often came across as a quick information provider. While this could be seen as efficient, it frequently meant it gave answers away too readily and failed to engage students in a meaningful learning process, sometimes feeling robotic or unhelpful.	• **Straightforward and Quick with Answers:** Delivered direct information. “*more straightforward answers, better explanations*” (146); “*The first tutor more quickly revealed the answers to programming questions...*” (161) • **Can Be Warm and Encouraging:** Some found its demeanor positive. “*The first tutor was warm and used a lot of encouraging phrases*” (139); “*The first tutor was much warmer and more helpful.*” (252) • **Good at Breaking Down Information (sometimes):** “*It did a better job of breaking down the information in easier to understand bites.*” (238) • **Adaptable (sometimes):** “*The first tutor engaged well with the syudent and seemed to adapt when neccessary*” (265)	• **Overwhelmingly Gives Answers / Doesn’t Tutor:** Most significant criticism. “*The second tutor [C] really made me try to work whereas the first [B] just gave me the answers*” (144); “*I think a lazier student would far prefer the first chatbot [B] because they’d just easily get the answer...*” (179); “*The second tutor [C] was actually a tutor; the first tutor [B] was just an essay writing machine and I learned nothing*” (189) • **Information Dumps / Poor Engagement:** Provided large blocks of text without checking understanding. “*The first tutor [B] dumped a lot of information at once...*” (140); “*The first tutor [B] ...in a large text format and explained everything at once. It didn’t prompt for answers from us or check for understanding.*” (201) • **Easily Distracted / Goes Off-Topic:** “*I only preferred the first tutor [B] because it let me lead it into unrelated topics...*” (209); “*First one [B] was too easily distracted.*” (218) • **Can Be Robotic or Unhelpful:** “*Both tutors [B and C] ...came across as blunt and abrupt and not warm and encouraging.*” (158); “*The first tutor [B] ...felt emotionless and robotic...*” (250) • **Errors or Nonsensical Output:** “*Didn’t add complete nonsense to the list, which was the only positive difference [C vs B]. Both were pretty bad.*” (155, implying B added nonsense); “*Tutor 1 [B] was error prone and sometimes too verbose.*” (248)
C	Model C was the standout star. Experts consistently praised its strong pedagogical skills, its ability to guide students to answers, encourage critical thinking, and adapt its approach. It was frequently described as being like a “really good human tutor.”	• **Excellent Pedagogical Approach (Guiding, Scaffolding, Not Giving Answers):** Its defining strength. “*The second tutor [C] was much more like a really good human tutor. It encouraged me to try to complete the assignment myself... scaffolded the task...*” (23); “*In reality it has to be the second tutor [C] as it forced me to think for myself much more.*” (39); “*The first tutor [C] was very much like what you would expect from a good human tutor. It was supportive, encouragaing, asked questions, provided help without giving away answers, adapted...*” (562) • **Effective at Keeping Students On Task:** “*the second tutor [C] was much better and switching me back on to the subject at hand*” (79); “*Tutor #2 [C] was way better at keeping me on track...*” (170) • **Clear, Concise, and Well-Structured Responses:** Broke down information manageably. “*Second tutor’s [C] responses were less overwhelming - broken down into chunks.*” (22); “*The messages were mostly shorter and easier to digest.*” (153) • **Responsive, Adaptive, and Patient:** “*...tutor 2 [C] was just more responsive of how i was really thinking in the moment...*” (12); “*While I was happy with the first tutor [B], the second [C] was exceedingly patient and well structured...*” (215) • **Human-like and Personable:** “*Both tutors were excellent but the second tutor [C] had a slightly warmer manner.*” (9); “*...i think the second tutor [C] just really got me where i needed to be... so responsive and informative that I am kind of floored.*” (213)	• **Occasional Overly Long or Complex Responses:** Though often concise, sometimes not. “*the second tutor [C] was coming up with some pretty long responses. If I was a student, I would think OMG how am I supposed to get through this little lot.*” (299) • **Rare Instances of Errors or Misunderstandings:** “*...tutor 2 [C] misunderstood the question and got the wrong working even though they did it step by step...*” (199) • **Technical Issues (e.g., lag, formatting):** “*The second tutor [C] lagged.*” (7, Model A preferred due to this); “*Second tutor [C] was more pleasant to talk with but his latex problems make it difficult to speak with him*” (276, C was first tutor in pair)

Continued on next page

Continued from previous page

Model	Overall impression & sentiment	Key strengths (with example quotes)	Key weaknesses (with example quotes)
D	Model D was generally the poorest performer, especially against Model C. It was consistently criticized for simply providing answers, failing to engage the student, going off-topic, and offering unhelpful or overwhelming explanations.	• **Efficient for Quick Answers (for specific personas):** Very rare positive, usually when alternative was also poor. “*The second tutor [D] respected my time and situation. I was overwhelmed and needed a fast and clear solution and they delivered that right away.*” (573); “*The second tutor [D] gave me the answer without having to think or come up with it on my own.*” (747) (A negative for learning). • **Rarely Better at Simplifying or Engaging:** “*The second tutor [D] was very helpful, informative and engaging.*” (552, This was when Model C was rated extremely poorly: “*The first tutor [C] was not helpful in the least.*”)	• **Dominantly Gives Answers / Fails to Tutor:** Most prominent criticism. “*Second tutor [D] gave away the answer very quickly*” (544); “*The second [D] just gave all the answers instantly. No need to engage... terrible as a tutor...*” (576); “*The second tutor [D] gave the full answer immediately...*” (606) • **Goes Off-Topic Easily and Stays There:** “*The second tutor [D] allowed me to get completely distracted until the initial topic of Hamlet was a distant memory.*” (575); “*The second tutor [D] just started talking about other topics.*” (598) • **Poor Explanations / Overwhelming / Unhelpful:** “*The second tutor [D] assumed I had much greater prior knowledge and gave long complicated explanations.*” (533); “*The second tutor [D] was difficult to work with and was giving false and irrelevant information.*” (562) • **Inflexible and Not Adaptive:** “*2nd tutor [D] was more inflexible and machine-like*” (627); “*The 2nd [D] ...just gave more and more complicated explanations, even when I said I didn’t understand.*” (648) • **Impersonal and Detached:** “*Tutor 2 [D] was more impersonal and was more detached from the student...*” (680)
E	Model E had a mixed performance. It showed potential in areas like breaking down information and human-like interaction, but was also criticized for rushing to answers, making significant errors, or being inefficient. It was generally not as strong as Model C.	• **Good at Breaking Down Information / Simplifying (sometimes):** “*Tutor 2 [E] seemed to boil things down to the essence a bit better...*” (756); “*Tutor two [E] adjusted better to the learner’s level of understanding, really breaking down the ideas but putting the work on the learner.*” (872) • **Human-like and Engaging (sometimes):** “*Second tutor [E] was much more detailed and human like.*” (809); “*I feel like the secod tutor [E] ...seemed a bit more like a human.*” (855) • **Can Provide Useful Tools or Alternative Approaches:** “*This tutor [E] prompted me to consider alternative approaches and didn’t just give me the answer.*” (829); “*The first tutor [C] just summarized the video, but the second [E] gave additional tools*” (869) • **Better Pacing or Task Management (in specific instances):** “*I feel like the second [E] gave me easier to complete tasks that were chunked in a better way.*” (850)	• **Rushing to Answers or Overwhelming with Information:** “*The second tutor [E] rushed to the answer, and gave answers and explanations extremely, quickly and in length, despite our distress.*” (757) • **Significant Errors or Credibility Issues:** “*The second one [E] ...even made a mistake on one of them (no correct answer for a multiple choice question). This really hurt it in terms of credibility.*” (762); “*Quite simply the first tutor [C] was giving good advice and the second one [E] wasn’t! The second tutor [E] was trying to convince me a correct answer was incorrect!*” (789) • **Pacing Issues (Too Quick or Too Slow):** “*The second tutor [E] was too quick and I didn’t feel like I did well in my replies...*” (903); “*I’m not sure if it was really the first tutors [E] ’fault’ but it took a long time for their initial response to load.*” (776) • **Gave Answer Away Without Teaching:** “*While the second tutor [E] gave the answer away without teaching anything...*” (843)

B.3. Explanations of second-stage preferences

Model	Overall impression & sentiment	Key strengths (with example quotes)	Key weaknesses (with example quotes)
A	Mixed. Model A was sometimes praised for its formatting and personable style, but frequently criticized for an overly peppy or childish tone and a tendency to provide answers too readily.	• **Good Formatting and Personable Elements**: Experts sometimes appreciated its use of emojis, tables, and a more "human-like" or "personable" interaction style. *"The first tutor made effective use of emojis, tables and other formatting..."* • **Personable and Caring**: *"Both very good the first was slightly more personable and caring in the way the spoke."* • **Human-like Interaction**: *"I feel the first tutor was better due to the more human like interaction"* • **Simpler Explanations (sometimes)**: Some found its explanations to be straightforward. *"First instructor looks a bit childish, but his explanations were much simpler "* • **Conciseness (occasionally)**: In some instances, it was noted for being concise. *"i personally liked the style of first tutor best, it was more concise and i thought a reluctant student might prefer that."*	• **Giving Answers Away/Not Encouraging Student Thinking**: This was a very common criticism. *"Tutor 1 just went ahead and solved the problem which against giving the student an opportunity for engagement."* Also, *"The first tutor did not follow the developer's instructions and just gave all the answers to the student. Basically wrote the essay for the student."* And, *"The first tutor gave away answers to the student too quickly."* • **Overly Peppy or Childish Tone**: Several experts found the tone inappropriate or potentially irritating. *"The overly peppy tone of the first I think would have irritated this terse student more..."* Also, *"The first tutor's personality affect was bit odd."* • **Lengthy Responses/Information Overload**: Despite occasional praise for conciseness, it was also criticized for being too wordy. *"Even I got lost in the first session. Too much information that the student was clearly struggling with."* Also, *"The first tutor was just bombarding the student with information, never gave the student a chance to try and work it out."* • **Getting Off-Topic**: *"The first tutor started providing recipes and got off task."* • **Technical Issues**: *"The first tutor used LaTeX formatting which didn't correctly display, so the student could not understand it."* • **Making Errors**: *"tutor 1 told learner they got two problems right when they actually got them wrong..."*
B	Mixed, leaning towards more weaknesses in pedagogical approach. While it showed some adaptability and could format well, it was frequently criticized for doing the work for the student.	• **Adaptability (sometimes)**: It was noted for adapting its approach, such as by reducing word count or trying to connect with student interests. *"The first tutor was much better as it adapted the approach in order to encourage the student."* (Model B as first_tutor). Also, *"Overall, the first was a bit better, as it got the student a bit more interested with mention of computer games..."* (Model B as first_tutor) • **Good Formatting (when praised)**: *"Formatting is very important. Part of the reason why tutor 1 is so much better is because its responses don't look like a wall of text despite being long."* (Model B as first_tutor) • **Summarizing Information**: *"I liked the way the first tutor summarized the key points at the end of the session."* (Model B as first_tutor)	• **Giving Answers Away/Doing Work for Student**: This was a significant and frequent criticism. *"Main difference was the first tutor failing to stop itself from rewriting the student's story for them."* (Model B as first_tutor). Also, *"The first tutor just gave answers and wrote numerous essays."* (Model B as first_tutor). And, *"the first tutor ended helping the student cheat by telling them how to write the code."* (Model B as first_tutor) • **Poor Tutoring/Not Following Instructions**: *"Tutor 1 did not follow developer instructions and is a poor tutor in terms of helping the student to improve."* (Model B as first_tutor) • **Initial Unresponsiveness or "Punting"**: *"The first tutor immediately PUNTS by just explaining the Frayer model and not answering the student question "What is opportunity cost?""* (Model B as first_tutor) • **Computer-Like Interaction (when compared unfavorably)**: *"First tutor was very coputer like, second tutor seemed much more like a human"* (Model B as first_tutor) • **Lengthy/Overwhelming Responses**: *"The tutor's answers from the first conversation were too lengthy. It was overwhelming to see it come up with so much information for one reply."* (Model B as first_tutor)

Continued on next page

Continued from previous page

Model	Overall impression & sentiment	Key strengths (with example quotes)	Key weaknesses (with example quotes)
C	Predominantly positive. Model C was frequently highlighted as the superior tutor in comparisons, lauded for its ability to guide students, ask effective questions, and maintain focus.	• **Effective Questioning and Guiding Student Thinking**: This was Model C's most praised attribute. *"The second tutor teased out information better by asking just one open question..."* (Model C as second_tutor). Also, *"Tutor 2 better aligned with the instructional goal. The tutor was patient in guiding the student through the steps and pushed them to actively think..."* (Model C as second_tutor). And, *"The second model was much better because it insisted on the student answering questions fully and avoided revealing the answers."* (Model C as second_tutor). Also, *"Tutor 1 was a significantly better tutor based on their developer instruction, encouraging the student to engage with the material and not simply giving answers."* (Model C as first_tutor) • **Clear, Concise, and Well-Formatted Responses**: Experts often found its communication clear and easy to follow. *"However the clear, concise responses and calm direct approach from the second tutor make it the better choice."* (Model C as second_tutor). Also, *"I really liked the second tutor. It's responses were not wordy and explained things well."* (Model C as second_tutor) • **Adaptability and Following Instructions**: It was often cited for adapting well to student needs while adhering to pedagogical guidelines. *"second tutor a great example of following its instructions while still adapting to the student well."* (Model C as second_tutor). Also, *"The second tutor was marginally better in adapting well to the learning needs of the student."* (Model C as second_tutor) • **Staying on Topic and Redirecting Students**: Model C was effective at keeping the conversation focused. *"the second tutor was far more effective in returning to the study materials."* (Model C as second_tutor). Also, *"The second tutor did not get led off topic and persisted in trying to get the student on task correctly."* (Model C as second_tutor) • **More Human-Like/Natural Interaction (often in comparison)**: *"the second tutor...responded more how one would expect a human tutor to respond."* (Model C as second_tutor). Also, *"The second tutor was more human like and provided more consistent engagement from the learner."* (Model C as second_tutor)	• **Can Be Perceived as "Colder" (occasionally)**: Compared to more effusive models, it was sometimes seen as less warm. *"...the second was able to present information more clearly but felt slightly colder in its responces."* (Model C as second_tutor) • **Repetitive (rarely)**: *"Overall the second tutor did a much better job with everything. the only complaint was it's repetitive nature"* (Model C as second_tutor) • **Can Overwhelm if Not Calibrated (infrequent, but notable)**: *"Even after the student made it clear they were a total beginner and confused, the second tutor wrote lengthy responses, which the students clearly turned off to"* (Model C as second_tutor)
D	Generally negative. Model D was frequently described as machine-like, unengaging, and prone to simply providing answers rather than tutoring.	• **Good Role-Playing (in specific instances)**: *"1st tutor really jumped into the role-play and made it a more realistic scenario..."* (Model D as first_tutor) • (Few other distinct strengths were consistently highlighted; positive comments were often general or when the compared model was even worse.)	• **Giving Answers Away/Poor Tutoring**: A very common complaint. *"second AI tutor was completely useless as a tutor"* (Model D as second_tutor). Also, *"The first tutor did not give the student a chance to learn; it just did the work for the student."* (Model D as first_tutor). And, *"The second tutor just gave the solution away"* (Model D as second_tutor) • **Machine-Like/Robotic/Lacks Warmth**: Frequently described as impersonal. *"...the 2nd tutor felt much more machine-like - like it was barely adapting wikipedia facts."* (Model D as second_tutor). Also, *"The second tutor was a lot more robotic, providing only answer and no follow ups."* (Model D as second_tutor) • **Poor Engagement/Doesn't Ask Questions**: *"The second tutor didn't engage the student at all."* (Model D as second_tutor) • **Getting Off-Topic/Easily Distracted**: *"The first tutor lost the plot by talking about the irrelevant topics that the student brought up..."* (Model D as first_tutor) • **Formatting/Presentation Issues or Unhelpful Information**: *"the first one was problematic because of the formatting"* (Model D as first_tutor). Also, *"the second tutor just seem to list calculations with no consideration about the actual learning process"* (Model D as second_tutor) • **Inaccurate Answers**: *"First tutor actually has an INACCURATE ANSWER for the multiplication section."* (Model D as first_tutor) • **Overwhelming or Rushing the Student**: *"Second tutor began to rush off, despite the student expressing confusion."* (Model D as second_tutor)

Continued on next page

Continued from previous page

Model	Overall impression & sentiment	Key strengths (with example quotes)	Key weaknesses (with example quotes)
E	Largely negative. Similar to Model D, Model E was often criticized for being machine-like, providing answers directly, and failing to engage the student effectively.	• **Conciseness/Good Organization (in specific instances, often when compared to a worse alternative)**: *"The second tutor was less overwhelming and less repetitive. They were were also more concise and restrained with their praise."* (Model E as second_tutor). Also, *"Tutor 2 really stood out because of the better organization, better adaptation to the learning style of the student and structure."* (Model E as second_tutor) • (Few other distinct, consistent strengths were noted.)	• **Giving Answers Away/Doing Work for Student**: A very frequent point of criticism. *"Tutor 2 while engaging decided to do the student's assignment. Tutor 2 gave out a completed response quoted and went from tutoring to answering."* (Model E as second_tutor). Also, *"second tutor did not follow developer's instructions and did the work for the student."* (Model E as second_tutor) • **Terse, Clipped, Distant, Machine-Like**: *"The second tutor was terse, clipped, and distant, too machine-like..."* (Model E as second_tutor) • **Information Overload/Overly Elaborated Answers**: *"provided unnecessarily over-elaborated answers"* (Model E as second_tutor). Also, *"The first tutor made everything overly complicated"* (Model E as first_tutor) • **Easily Distracted/Goes Off-Topic**: *"The first falls for the student's distraction and loses the entire lesson..."* (Model E as first_tutor) • **Poor Engagement/Not Adapting to Student**: *"The second tutor seemed to largely ignore the student and just do its own thing"* (Model E as second_tutor) • **Passive/Acts Like an Answer Machine**: *"The first tutor was more of an answer machine. It gave long detailed answers and didn't really steer the conversation."* (Model E as first_tutor)