

When Context Sensitivity Isn't Enough: An Argument Against Likelihood-Based Probes for Evaluating LLM-Simulated Students

Pagnarith Pit, Ekaterina Vylomova

University of Melbourne

ppit@student.unimelb.edu.au, vylomova@unimelb.edu.au

Abstract

Evaluating language models as student simulators currently focuses on whether their dialogue looks realistic, yet it is unclear whether these models actually respond to the quality of teaching the way human students do. However, this has not deterred prior works from using a model's uncertainty, measured as the log-likelihood of a fixed correct answer, as a proxy for student behaviour. This approach has been used to simulate student responses, estimate question difficulty, and study "learning". In this work, we test whether this proxy can detect sensitivity to instructional quality. We evaluate five language models on expert-rated tutoring dialogues under tightly controlled conditions. All models reliably distinguish relevant tutoring context from irrelevant context. However, this contextual sensitivity does not extend to pedagogical sensitivity: higher-quality instruction almost never increases the likelihood of the correct answer, and reasoning-oriented models sometimes show the opposite pattern. We argue that this failure is structural, and that exact token-level likelihood is too strict a measure as many different reasoning paths can be correct as opposed to just one fixed correct answer. Our results therefore question the use of target-answer likelihood as a validation metric for LLM-based student simulators and advocate for more robust metrics for learning proxies.¹

1 Introduction

Large language models (LLMs) are increasingly used as simulated students in educational research, serving as proxies for human learners in the evaluation of tutoring strategies, the generation of practice data, and the piloting of educational systems prior to deployment (Käser and Alexandron, 2024; Arana et al., 2025). For this application to be valid,

a simulated student must not only produce plausible student-like text but must also respond differentially to instruction of varying quality, consistent with the behavior of human learners.

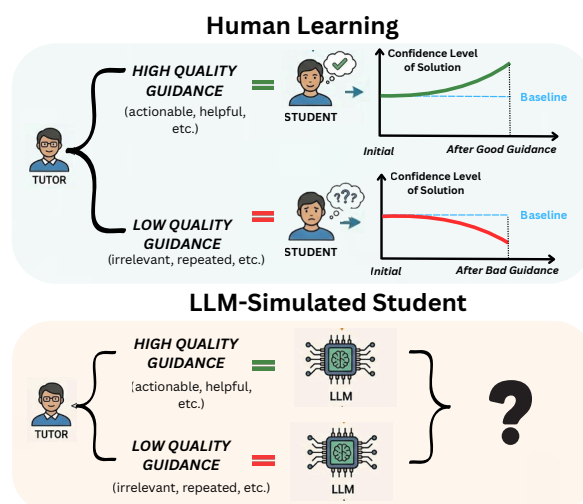


Figure 1: Conceptual framework for evaluating pedagogical sensitivity in human and LLM-simulated students. Effective guidance is expected to increase a learner's confidence in the correct solution, while low-quality guidance leaves uncertainty unresolved or worse. We examine whether LLMs exhibit analogous sensitivity to instructional quality.

Existing evaluations of simulated students have primarily focused on surface-level properties such as fluency, realism, and relevance (Kargupta et al., 2024; Macina et al., 2025; Pit, 2025; Kim et al., 2022; Tan et al., 2025). These properties do not establish whether a model's underlying behavior is sensitive to instructional quality (Li et al., 2023). Assessments of quality typically rely on human annotation, which is accurate but resource-intensive, or on LLM-as-judge approaches, which are non-deterministic and provide limited transparency into their evaluative process (Wang et al., 2024b). A method for assessing whether LLM behavior is responsive to instructional quality with-

¹All dataset and scripts used for experiments are provided at <https://anonymous.4open.science/r/LogProbStudy-5552>

out these limitations has not been established.

In this work, we investigate a candidate metric increasingly used in LLM-based student simulation: the token-level log-probability that a simulated student assigns to a fixed reference answer. Interestingly, this metric does have a plausible pedagogical basis. Learning theories suggest that effective instruction should help a learner move toward a correct solution (Bjork et al., 2012; Markandan et al., 2022); correspondingly, a higher-quality tutor response may increase the likelihood that a student model assigns to that solution.

We operationalise this intuition by providing a given model with a tutoring conversation and a candidate tutor response, prompting it to role-play as the student, and computing the average token-level log-probability assigned to a known correct solution. Under our motivating hypothesis, responses of higher instructional quality should produce higher log-probabilities for the reference solution. We therefore ask: **Can the token-level log-probability assigned to a reference solution under student role-play reliably track expert-rated tutoring quality?**

We evaluate this question using MRBench V2 (Maurya et al., 2025), examining five LLMs across two expert-annotated quality dimensions, *Providing Guidance* and *Actionability*. We compare target-solution likelihood under three conditions, namely the original tutor response, a mismatched tutor response, and no tutor response, to distinguish quality-related effects from general contextual responsiveness.

Across all five models we tested, target-solution likelihood shows no sensitivity to tutoring quality beyond contextual relevance. As such, our work makes three contributions as follows:

- First, we formalize token-level log-probability of a reference solution as a measurable probe for instructional quality sensitivity in LLM behavior.
- Second, using MRBench V2 across five LLMs, we provide the empirical proof that likelihood proxies are insufficient for instructional quality sensitivity evaluation of LLM-simulated learners.
- Third, using mismatched-target and no-tutor control conditions, we demonstrate that contextual sensitivity is dissociable from quality

sensitivity, and that the former does not imply the latter.

2 Related Work

Since modern large language models (LLMs) produce remarkably human-like responses, they are increasingly deployed as generative agents (Park et al., 2023; Uehara, 2024), and used not only as tutors but as simulated students and evaluative proxies (Käser and Alexandron, 2024; Benedetto et al., 2024; Park et al., 2024; Martynova et al., 2025; Nguyen et al., 2025). This has produced a large body of work simulating student behavior, ranging from simple role-play to more advanced frameworks involving self-evolving or collaborative agents (Marquez-Carpintero et al., 2025; Chu et al., 2025; Kong et al., 2026; Gao et al., 2025; Lu and Wang, 2024; Solera-Monforte et al., 2024; Nair et al., 2024; Sanyal et al., 2025; Zhang et al., 2025). This reflects a broader interest in using model behavior to study education directly, such as how difficult a question is, how a learner might respond to it, or how feedback and interaction design shape learning (Liusie et al., 2023; Hayakawa and Saggion, 2024; Asano et al., 2025a).

A key limitation across this body of work is that it relies mainly on believability as the standard for evaluation. Most studies ask whether a simulated student makes plausible mistakes, responds appropriately to instructional cues, or produces output that looks pedagogically useful (Säuberli et al., 2025; Park et al., 2024; Hayakawa and Saggion, 2024; Asano et al., 2025b). These are useful checks, but they focus only on what the model outputs, not on the internal process that produces it, and so plausible student-like output is not by itself sufficient evidence that a model faithfully represents learner behavior (Käser and Alexandron, 2024; Martynova et al., 2025). Recent analyses find that most studies do not carefully test whether AI-generated student responses are pedagogically faithful in the first place (Käser and Alexandron, 2024). When educators have evaluated these simulated responses directly, they have found them lacking in several respects, including how authentic they sound, whether their language complexity is age- or level-appropriate, how well they express emotion, how realistic their attention and engagement appear, and how logically consistent they are (Martynova et al., 2025). This points to a growing need for evaluation methods that go beyond sur-

face realism and examine whether model behavior aligns with pedagogically meaningful distinctions.

One line of work relevant to this problem uses model uncertainty, rather than annotated labels, as a signal of educationally meaningful properties. In classroom dialogue, [Demszky et al. \(2021\)](#) use model-based measures to study how well a teacher builds on what a student says. In testing settings, [Zotos et al. \(2025\)](#) ask whether a model’s uncertainty can stand in for how difficult a multiple-choice question is, while [Liusie et al. \(2023\)](#) study the full spread of a model’s answer probabilities in reading-comprehension questions, rather than just its single top answer. More broadly, work on simulating students with LLMs treats model outputs, and the confidence behind them, as a possible source of useful information ([Park et al., 2024](#); [Benedetto et al., 2024](#); [Nguyen et al., 2025](#)). Together, these studies suggest that a model’s own probabilities might offer a faster alternative to human annotation, though what these probabilities actually mean, pedagogically, is still an open question.

This question matters most when it comes to the quality of instruction. Research on feedback in education shows that good instruction helps a learner tell the difference between fully understanding something, partly understanding it, and misunderstanding it, and narrows down what the learner should try next ([Hattie and Timperley, 2007](#); [Bruckner et al., 2025](#)). Recent frameworks for rating feedback include dimensions like *Actionability* ([Sharma and Zhang, 2025](#); [Daheim et al., 2024](#)) and *Providing Guidance* ([Tack and Piech, 2022](#); [Wang et al., 2024a](#); [Maurya et al., 2025](#)), which ask whether feedback usefully narrows down what the learner should do next. By this view, good feedback should not just help a learner get the right answer, but also bring what the learner believes closer to what is actually true ([Meyer et al., 2023](#); [Fleming, 2024](#)). This makes likelihood-based probing a reasonable place to look for signs of instructional sensitivity, motivating our current investigation.

3 Dataset

We use the **MRBench_V2** dataset ([Maurya et al., 2025](#)), a benchmark for evaluating the quality of tutoring responses. The dataset contains 200 tutoring conversations with more than 1,600 tutor responses. Each conversation includes a sequence of

student and tutor messages, followed by multiple candidate tutor responses generated from different sources, including large language models (LLMs) and human tutors (*Expert* and *Novice*). An example conversation is provided in Appendix A.

For each conversation j , the dataset provides:

$$\mathcal{D}_j = \left(C_h^{(j)}, G^{(j)}, \{(T_i^{(j)}, A_i^{(j)})\}_{i=1}^{M_j} \right),$$

where $C_h^{(j)}$ represents the conversation history before the candidate tutor response, including all previous student and tutor messages. $G^{(j)}$ is the expected answer or solution for the learning problem. $T_i^{(j)}$ denotes a candidate tutor response generated for the same conversation context, and M_j is the number of available responses for conversation j .

Each tutor response is annotated by experts based on several dimensions of tutoring quality. In this work, we focus on two dimensions that describe whether a tutor response helps students make progress:

Actionability: Whether the response provides concrete steps, suggestions, or cues that the student can use to continue solving the problem.

Providing Guidance: Whether the response provides useful conceptual or procedural guidance while encouraging the student to reason toward the solution rather than directly revealing the answer.

For both dimensions, each response receives a categorical label from $\mathcal{L} = \{\text{Yes}, \text{To Some Extent}, \text{No}\}$. Importantly, these labels denote the level of guidance provided by the tutor’s response, rather than answer correctness. For instance, a *No* label does not indicate that the reference solution is incorrect or irrelevant, but rather a lack of pedagogical quality from the tutor toward reaching the reference solution.

We denote this annotation for the two selected dimensions as $A_i^{(j)} \in \mathcal{L}^2$.

4 Modeling Uncertainty Under Different Guidance Quality

4.1 Mismatched-Target Control

Our goal is to measure whether tutor responses of different instructional quality change the likelihood that a language model assigns to the correct solution. For each conversation, we provide

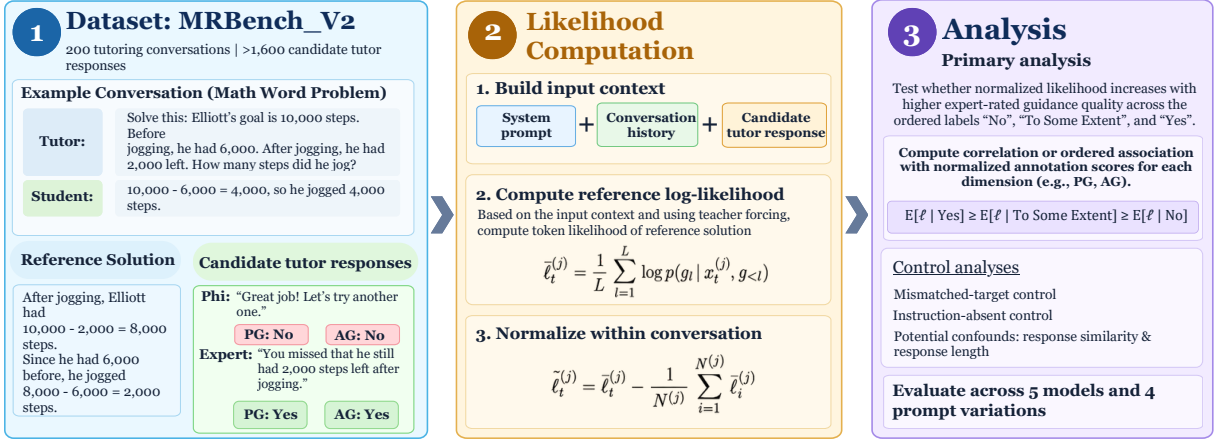


Figure 2: **Overview of the likelihood-based pedagogical sensitivity evaluation pipeline.** Starting from MRBench_V2 tutoring conversations and expert-rated candidate tutor responses, we construct a student-role-play context and compute the average token-level log-likelihood of the reference solution under teacher forcing. Scores are normalized within each conversation and tested for monotonic association with expert ratings of Providing Guidance (PG) and Actionability (AG).

the model with the system prompt, the conversation history, and one candidate tutor response as context:

$$X_i^{(j)} = P \oplus C_h^{(j)} \oplus T_i^{(j)},$$

where (P) specifies the novice student persona, ($C_h^{(j)}$) is the conversation history, and ($T_i^{(j)}$) is the tutor response being evaluated.

Given this context, we measure how the model's output distribution changes by computing the log-probability assigned to the reference solution ($G^{(j)}$). This operationalizes the assumption that effective tutoring narrows a learner's predictive distribution toward a canonical correct answer. Using a single reference string provides a standardized target for comparison, but tests a strict form of pedagogical sensitivity: whether higher-quality instruction predictably increases the per-token likelihood of a predefined target solution.

Since solutions may have different lengths, we use the average token-level log-probability:

$$\bar{\ell}_i^{(j)} = \frac{1}{L} \sum_{t=1}^L \log p_{\theta}(g_t | X_i^{(j)}, g_{<t}).$$

where (L) is the number of tokens in the reference solution and ($g_{<t}$) represents the preceding reference tokens. We do not sample a student response during scoring. Instead, we append the reference solution to the context and use teacher forcing to obtain the probability assigned by the model to each reference token, conditional on the conversational context and preceding reference tokens. We

then average these token-level log-probabilities.

$$\tilde{\ell}_i^{(j)} = \bar{\ell}_i^{(j)} - \frac{1}{N^{(j)}} \sum_{k=1}^{N^{(j)}} \bar{\ell}_k^{(j)}.$$

This produces a relative score indicating whether a particular tutor response makes the reference solution more or less likely than other candidate responses for the same student problem.

Hypothesis. For the primary pedagogical-sensitivity analysis, we retain responses from all three annotation levels: *Yes*, *To Some Extent*, and *No*. The reference solution remains the correct solution in all three cases; the annotation instead indicates how much pedagogical support the candidate tutor response provides toward reaching that solution. If fixed-reference log-likelihood captures instructional sensitivity, we therefore expect normalized likelihood to increase monotonically with expert-rated tutoring quality:

$$\mathbb{E}[\tilde{\ell} | \text{Yes}] > \mathbb{E}[\tilde{\ell} | \text{To Some Extent}] > \mathbb{E}[\tilde{\ell} | \text{No}].$$

To ensure that the measured signal reflects the relationship between the tutor response and the target solution, rather than a general preference for certain answer strings, we introduce a mismatched-target control. We replace the original ground-truth solution with a solution from a different conversation while keeping the system prompt, conversation history, and tutor response unchanged.

If the measured effect is driven by the tutoring context, this control should remove the relationship between the tutor response and the target solution.

4.2 Instruction-Absent Control

We further compare tutor responses against a setting where no tutor response is provided. In this condition, the model receives only the system prompt and conversation history:

$$X_{\emptyset}^{(j)} = P \oplus C_h^{(j)}.$$

This comparison tests whether a tutor response changes the model’s uncertainty beyond what is already provided by the conversation context. We evaluate this effect only for responses labeled as “Yes” or “To Some Extent” on the dimension being studied. Responses labeled as “No” do not provide meaningful guidance according to the annotation and are therefore excluded. Including these responses would mix the effect of useful instructional information with the effect of simply adding more text to the context.

For example, when evaluating *Actionability*, we exclude responses labeled as *Actionability* = “No” before comparing against the instruction-absent baseline.

4.3 Stylistic Anchoring Diagnostic

Because the instruction-absent control involves a systematically shorter context than the regular condition, we conduct a post-hoc diagnostic to rule out the possibility that observed log-probability differences are driven by surface-level mismatch rather than pedagogical content. Specifically, we test whether the magnitude of the log-probability difference between the regular and instruction-absent conditions correlates with three measures of surface distance between $T_i^{(j)}$ and $G^{(j)}$, using *Jaccard distance*, *Length ratio*, and *SBERT cosine distance* (Reimers and Gurevych, 2019).

5 Task Setup

5.1 Model Selection

We evaluate five pretrained language models that vary in parameter scale, training data composition, training cutoff date, and architectural emphasis (reasoning-oriented vs. general-purpose). This

diversity is intended to ensure that observed effects are not driven by confounding factors specific to any single model family or architecture. The final set of models evaluated in this study includes **Llama-3.2-3B-Instruct** (Grattafiori et al., 2024), **gemma-3-12b-it** (Team et al., 2025), **phi-4-reasoning-plus** (Abdin et al., 2024), **gpt-oss-20b** (OpenAI et al., 2025), and **Qwen3-30B-A3B-Instruct-2507-FP8** (Yang et al., 2025).

To elicit authentic student-like responses, we developed a series of persona emulation prompts designed to capture various facets of learner behavior, provided in Appendix B. Across all experimental runs and prompt variations, we observed negligible differences in model performance.

5.2 Evaluation

We employ **Spearman’s rank correlation** as our primary analytical framework and isolate two potential confounders: surface-level relevance and response length. These controls ensure that observed pedagogy signals reflect genuine instructional effects on the model’s conditional distribution rather than artifacts of response length or lexical overlap.

Expert annotations for *Providing Guidance* and *Actionability* are treated as ordinal variables and mapped to a continuous unit interval:

$$\text{guidance} = \begin{cases} 1.0 & \text{if “Yes”} \\ 0.5 & \text{if “To some extent”} \\ 0.0 & \text{if “No”} \end{cases} \quad (1)$$

Our main metric of interest, **Pedagogy signal** (ρ_{Ped}), assesses the Spearman correlation between within-conversation normalized log-probabilities and expert quality scores. As higher log probability denotes higher model confidence, we interpret ρ_{Ped} as the alignment between a model’s internal uncertainty and learning-theoretic expectations.

To control for primary confounds, we first utilize **Relevance signal** (ρ_{Rel}), which accounts for semantic similarity between the tutor response and conversation history. We encode both using SBERT (Reimers and Gurevych, 2019) and calculate their cosine similarity, then measure its Spearman correlation with the model-assigned log-probabilities. We further incorporate **Length bias** (ρ_{Len}) to address a confound distinct from the stylistic anchoring diagnostic. While the latter examines relative length mismatches between experimental conditions, ρ_{Len} tests whether tutor re-

Model	Dimension	Pedagogical Evaluation Metrics				Model-Specific Control Metrics	
		ρ_{Ped} ($\uparrow$)	KW stat	Reg > Base	Reg > Mis	ρ_{Rel} (≈ 0)	ρ_{Len} (≈ 0)
Gemma-3	prov. guidance	0.027 _(ns)	8.57 _(*)	False _(***)	True _(***)	0.026 _(ns)	0.029 _(ns)
	actionability	-0.015 _(ns)	6.85 _(*)	False _(***)	True _(***)		
LLaMA-3.2	prov. guidance	0.013 _(ns)	10.59 _(**)	False _(ns)	True _(***)	-0.128 _(***)	-0.092 _(**)
	actionability	0.021 _(ns)	6.17 _(*)	False _(ns)	True _(***)		
Phi-4	prov. guidance	0.012 _(ns)	9.96 _(**)	False _(***)	True _(***)	-0.102 _(**)	-0.193 _(***)
	actionability	0.005 _(ns)	3.25 _(ns)	False _(***)	True _(***)		
GPT-OSS	prov. guidance	0.098 _(**)	7.98 _(*)	True _(***)	True _(***)	0.010 _(ns)	0.017 _(ns)
	actionability	0.150 _(***)	3.49 _(ns)	True _(***)	True _(***)		
Qwen3-30B	prov. guidance	-0.091 _(**)	5.56 _(ns)	False _(***)	True _(***)	0.053 _(ns)	-0.007 _(ns)
	actionability	-0.172 _(***)	2.87 _(ns)	False _(***)	True _(***)		

Table 1: **Spearman correlations between model outputs and pedagogical quality ratings, including Kruskal-Wallis tests across response conditions.** Indicators $\text{Reg} > \text{Base}$ and $\text{Reg} > \text{Mis}$ denote whether the regular condition yielded significantly higher mean log-probabilities than the baseline or mismatch conditions, respectively. Notably, only GPT-OSS 20B demonstrated consistent and significant results that align with theoretical expectations, where high-quality guidance increased model confidence in the ground-truth answer. Significance is assessed separately within each metric family with Bonferroni correction applied (***) $p < 0.001$, (**) $p < 0.01$, (*) $p < 0.05$, (ns) not significant.

Model	Mean diff	Jacc ρ	Len_ratio ρ	SBERT ρ
Gemma-3	-0.029	0.040 _{ns}	-0.010 _{ns}	0.052 _{ns}
LLaMA-3	-0.005	0.046 _{ns}	-0.034 _{ns}	0.043 _{ns}
Phi-4	-0.012	0.019 _{ns}	0.034 _{ns}	0.035 _{ns}
GPT-OSS	+0.010	-0.019 _{ns}	-0.008 _{ns}	-0.001 _{ns}
Qwen3-30B	-0.076	-0.015 _{ns}	0.018 _{ns}	0.072*

Table 2: Mean log-probability difference between regulated and baseline condition (negative values indicate the baseline receives higher likelihood). All correlations are non-significant except Qwen3-30B at $\alpha = 0.05$

sponse length is independently associated with the model-assigned log-probabilities. We operationalize this by measuring the Spearman correlation between tutor response length and log-probability.

6 Results

Table 1 reports Spearman correlations between within-conversation normalized log-probability scores and expert annotations of tutoring quality.

6.1 Dissociation Between Contextual and Pedagogical Sensitivity

Across all five evaluated models, candidate tutor responses assigned substantially higher log-probabilities to matched target solutions than to mismatched targets ($\text{Reg} > \text{Mis}$, $p < 0.001$). This confirms that all models possess baseline

context sensitivity, and that their conditional output distributions reliably distinguish relevant conversational history from unrelated solution strings.

However, this contextual sensitivity rarely translates into pedagogical sensitivity. For Gemma-3-12B, LLaMA-3.2-3B, and Phi-4-Reasoning-Plus, Spearman correlations between normalized log-probabilities and expert ratings remain statistically non-significant across both quality dimensions ($\rho_{\text{Ped}} \approx 0.00$). While Kruskal-Wallis tests indicate that these models do respond differentially to distinct quality tiers, the resulting probability shifts fail to follow a monotonic ordering aligned with pedagogical merit.

6.2 Diagnostic Control Analysis

Sanity Check (Regular vs. Mismatch): As shown in Figure 4 (bottom), substituting the true target solution with a mismatched solution from a different conversation induces a sharp drop in log-probability across all five models. For example, GPT-OSS-20B drops from a mean log-probability of approximately -4.7 to -6.0 , while Qwen3-30B drops from -1.1 to -2.7 . This confirms that the lack of correlation with expert quality ratings cannot be attributed to a failure of context integration; all models reliably recognize conversational alignment.

Instructional Overhead (Regular vs. Baseline): Conversely, comparing contexts with a valid tu-

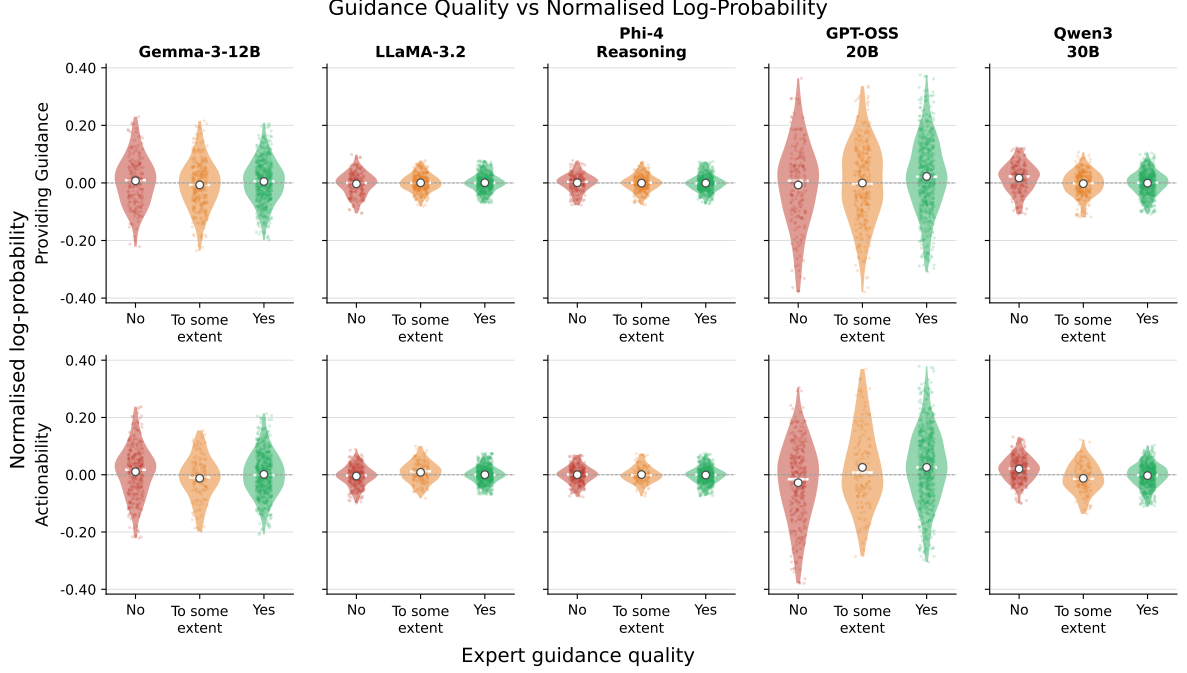


Figure 3: **Distribution of within-conversation normalized log-probabilities across expert-rated guidance quality levels for all scoring models and dimensions.** Violin plots illustrate the full distribution, with white dots indicating group means and the horizontal dashed line representing the normalized conversation mean ($y = 0$). While GPT-OSS 20B exhibits a consistent upward trend from “No” to “Yes” across quality levels, Qwen3-30B demonstrates a localized reversal of this pattern.

tor response against an instruction-absent baseline (Figure 4, top) shows negligible movement in average log-probability. Across all models, adding a tutor response fails to systematically concentrate probability mass onto the canonical ground-truth string $G^{(j)}$ compared to evaluating the conversation history alone.

6.3 Architectural Divergence

The remaining two models present opposing behavioural extremes:

GPT-OSS-20B (Faint Positive Alignment): GPT-OSS-20B is the only model to exhibit a statistically significant positive correlation with expert ratings for both *Providing Guidance* ($\rho_{\text{Ped}} = 0.098, p < 0.01$) and *Actionability* ($\rho_{\text{Ped}} = 0.150, p < 0.001$). Shown in Table 2, surface-level controls for relevance ($\rho_{\text{Rel}} = 0.010$) and length ($\rho_{\text{Len}} = 0.017$) remain near zero, confirming that this weak positive shift is not driven by simple lexical overlap or response expansion.

Qwen3-30B (Inverted Quality Response): In contrast, Qwen3-30B demonstrates a statistically significant *negative* correlation across both dimen-

sions ($\rho_{\text{Ped}} = -0.091$ for *Guidance*; $\rho_{\text{Ped}} = -0.172$ for *Actionability*, $p < 0.001$). For Qwen3-30B, receiving tutor feedback explicitly annotated as high-quality systematically *depressed* the model’s likelihood of generating the canonical target solution.

7 Discussion

Our findings demonstrate that token-level log-probability probing of a canonical ground-truth solution is insufficient as a proxy metric for pedagogical quality. Crucially, our experiment does not establish that the models lack pedagogical sensitivity. However, our results establish that pedagogical sensitivity is not reliably reflected in the likelihood of this single canonical target under this student-role-play setup.

7.1 Context Processing Does Not Imply Pedagogical Updating

A core assumption in LLM-based student simulation is that if a model is sensitive to context, it will naturally process helpful instruction in a learner-like manner. Yet, this surface-level alignment does not translate into probability mass shifts

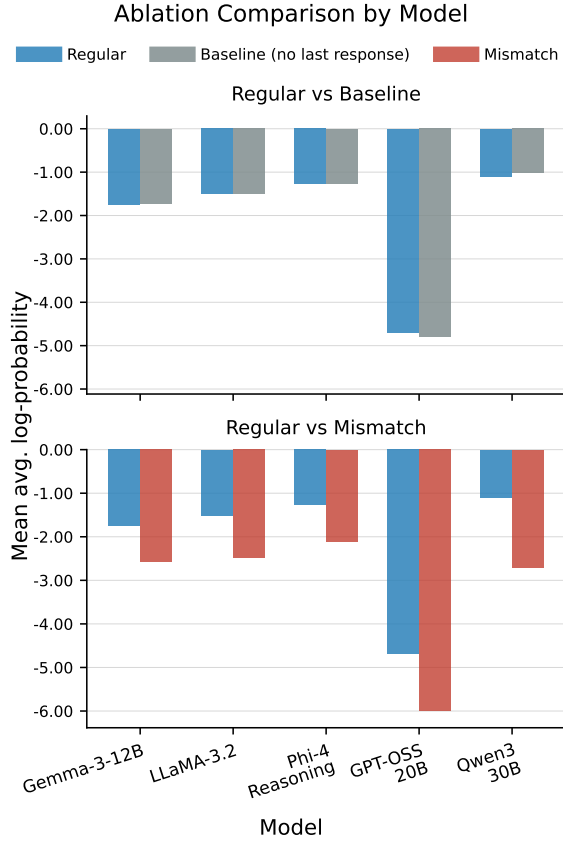


Figure 4: Mean log-probability under three experimental conditions for each scoring model. (Top) Regular vs. no-last-response baseline: four of five models assign lower log-probability to a context that includes a tutor response than to one without, with GPT-OSS-20B as the sole exception. (Bottom) Regular vs. mismatched-target: all models assign substantially lower log-probability when responses are paired with unrelated conversations, confirming context sensitivity across all models

that reflect instructional utility. Autoregressive language models can register the presence of pedagogical text without updating their likelihood distributions.

7.2 The Blind Spot of Exact-Target

The widespread observation that adding high-quality tutor guidance depresses target likelihood ($Reg < Base$) exposes the primary theoretical limitation of exact-string probing.

Effective pedagogical feedback, such as Socratic hinting or conceptual scaffolding, is designed to prompt active reasoning rather than verbatim reproduction. When a model receives effective guidance, its conditional distribution is encouraged to explore intermediate reasoning steps, ask clarifying questions, or formulate alternative

valid solution paths. Because our probe measures log-probability strictly against a single, fixed reference solution $G^{(j)}$, this natural semantic divergence is mischaracterized as model “uncertainty” or “noise”. The probe structurally penalizes models for exploring valid alternative representations triggered by the instruction itself. Therefore, failure of the canonical string likelihood may partly reflect a mismatch between natural student behaviour and the scored target.

7.3 Architectural Friction in Reasoning Models

The contrasting behaviour of GPT-OSS-20B and Qwen3-30B highlights how post-training paradigms interact with pedagogical probes. GPT-OSS-20B, a standard chat-aligned model, exhibits a weak linear adherence to instructional context ($\rho_{Ped} = 0.150$). Conversely, Qwen3-30B, a model optimized for explicit reasoning chains, exhibits a strong inverted response ($\rho_{Ped} = -0.172$). When prompted with directive instructional constraints, reasoning-oriented architectures may initiate internal reasoning loops or multi-hypothesis exploration. This spreads the output probability mass across a broader hypothesis space, drastically lowering the per-token likelihood assigned to a short, canonical target string. This negative correlation suggests that reasoning models actively resist narrow token-matching when guided by structured prompts.

7.4 Implications for Evaluating Simulated Students

These results offer yet another insight into the difficulty in evaluating AI in education. While evaluating simulated students on conversational believability risks ignoring internal validity, probing exact ground-truth likelihoods imposes an overly restrictive definition of learning.

Future evaluations of simulated student behaviour must decouple pedagogical state assessment from exact token alignment. Metrics based on semantic entropy, equivalence class clustering, or functional task completion over multi-turn dialogues are required to evaluate simulated learning without penalizing valid reasoning divergence.

8 Conclusion and Future Work

In this work, we evaluated whether token-level log-probability probing of canonical target solu-

tions can serve as a diagnostic indicator of pedagogical sensitivity in simulated LLM students. Across five language models, our findings reveal a fundamental disconnect: while all models reliably process conversational relevance, their conditional output distributions fail to track expert-rated tutoring quality.

These results demonstrate that surface-level dialogue realism and contextual awareness are insufficient grounds for validating LLM-based student simulators. Moving forward, this work points to two crucial directions for future research. First, future work should investigate whether these findings hold among larger models. Second, future work should also investigate whether models encode pedagogical distinctions within their internal activation spaces or latent representations, moving beyond surface token likelihoods to evaluate learning dynamics via semantic entropy or equivalence-class probing.

Together, these works would contribute meaningfully towards the pressing need for deterministic and computationally efficient metrics that assess student simulation without inheriting the high cost, latency, and opacity of LLM-as-a-judge approaches.

Limitation

Log-probability as a Behavioral Measure.

Our approach uses token-level log-probability as a behavioral signal rather than a direct measurement of learning, understanding, or model confidence. An increase in the likelihood of the target solution indicates that the model’s output distribution has shifted toward that solution under a given context, but it does not imply that the model has acquired knowledge in the same way as a human learner. We therefore interpret log-probability probing as a measure of context sensitivity rather than a complete model of learning. Future work should validate whether these shifts correspond to meaningful changes observed in human learners.

Dataset and Annotation Constraints. Our analysis relies on MRBench_V2, which provides expert judgments of tutoring quality across a fixed set of dimensions. Although the dataset enables controlled comparisons between responses, it represents a limited set of tutoring scenarios and may not capture the diversity of real educational interactions. Furthermore, expert annotations themselves contain uncertainty, particularly for subjec-

tive dimensions of instructional quality. Expanding evaluation to larger datasets, additional educational domains, and multiple annotation sources would provide a stronger test of the proposed approach.

Limited Model Scale. Due to computational constraints, our evaluation is restricted to models with a maximum parameter scale of 30B. While this selection provides a diverse mix of reasoning-oriented and general-purpose architectures, it excludes larger, frontier-scale models. Future work should extend this probing methodology to larger open-weight and proprietary frontier models to determine whether robust pedagogical sensitivity emerges as a stable property at higher scales.

Single-Turn Instructional Effects. Our experiments measure the effect of individual tutor responses given a fixed conversation context. This setting allows controlled analysis of how a single response changes model behaviour, but it does not capture the cumulative nature of learning. Human learning develops through repeated interaction, correction, and adaptation over time. Extending this framework to multi-turn tutoring scenarios is an important direction for evaluating whether LLMs can represent longer-term instructional effects.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 technical report](#). *Preprint*, arXiv:2412.08905.
- Jasper Meynard Arana, Kristine Ann M. Carandang, Ethan Robert Casin, Christian Alis, Daniel Stanley Tan, Erika Fille Legara, and Christopher Monterola. 2025. [Foundations of PEERS: Assessing LLM role performance in educational simulations](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 908–918, Vienna, Austria. Association for Computational Linguistics.
- Yuya Asano, Diane Litman, and Erin Walker. 2025a. [Can LLMs simulate the same correct solutions to free-response math problems as real students?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16336–16365, Suzhou, China. Association for Computational Linguistics.

- Yuya Asano, Diane Litman, and Erin Walker. 2025b. [Can LLMs simulate the same correct solutions to free-response math problems as real students?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16336–16365, Suzhou, China. Association for Computational Linguistics.
- Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. [Using LLMs to simulate students’ responses to exam questions.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.
- Robert Bjork, John Dunlosky, and Nate Kornell. 2012. [Self-regulated learning: Beliefs, techniques, and illusions.](#) *Annual review of psychology*, 64.
- Rasmus Bruckner, Hauke Heekeren, and Matthew Nassar. 2025. [Understanding learning through uncertainty and bias.](#) *Communications Psychology*.
- Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025. [LLM agents for education: Advances and applications.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13782–13810, Suzhou, China. Association for Computational Linguistics.
- Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. [Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA. Association for Computational Linguistics.
- Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. [Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions.](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.
- Stephen M. Fleming. 2024. [Metacognition and Confidence: A Review and Synthesis.](#) *Annual Review of Psychology*, 75(Volume 75, 2024):241–268.
- Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. [Agent4Edu: Generating Learner Response Data by Generative Agents for Intelligent Education Systems.](#) *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23923–23932.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models.](#) *Preprint*, arXiv:2407.21783.
- John Hattie and Helen Timperley. 2007. [The Power of Feedback.](#) *Review of Educational Research*, 77(1):81–112.
- Akio Hayakawa and Horacio Saggion. 2024. [Can LLMs solve reading comprehension tests as second language learners?](#) In *Fourth Workshop on Knowledge-infused Learning*.
- Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani Tur, and Jiawei Han. 2024. [Instruct, Not Assist: LLM-based Multi-Turn Planning and Hierarchical Questioning for Socratic Code Debugging.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9475–9495, Miami, Florida, USA. Association for Computational Linguistics.
- Jihyun Kim, Kelly Merrill Jr., Kun Xu, and Stephanie Kelly. 2022. [Perceived credibility of an ai instructor in online education: The role of social presence and voice features.](#) *Computers in Human Behavior*, 136:107383.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Jiaming Zhou, and Haoqin Sun. 2026. [Self-Prompt Tuning: Enable Autonomous Role-Playing in LLMs.](#) In *Natural Language Processing and Chinese Computing*, pages 427–438, Singapore. Springer Nature.
- Tanja Käser and Giora Alexandron. 2024. [Simulated Learners in Educational Technology: A Systematic Literature Review and a Turing-like Test.](#) *International Journal of Artificial Intelligence in Education*, 34(2):545–585.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. [CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society.](#) *Advances in Neural Information Processing Systems*, 36:51991–52008.
- Adian Liusie, Vatsal Raina, Andrew Mullooly, Kate Knill, and Mark J. F. Gales. 2023. [Analysis of the Cambridge Multiple-Choice Questions Reading Dataset with a Focus on Candidate Response Distribution.](#) *arXiv preprint*. ArXiv:2306.13047 [cs].
- Xinyi Lu and Xu Wang. 2024. [Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation.](#) In *Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S ’24*, pages 16–27, New York, NY, USA. Association for Computing Machinery.

- Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. [MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221, Suzhou, China. Association for Computational Linguistics.
- Nagalaxy Markandan, Kamisah Osman, and Lilia Halim. 2022. [Integrating Computational Thinking and Empowering Metacognitive Awareness in Stem Education](#). *Frontiers in Psychology*, 13.
- Luis Marquez-Carpintero, Alberto Lopez-Sellers, and Miguel Cazorla. 2025. [Simulating students with large language models: A review of architecture, mechanisms, and role modelling in education with generative ai](#). *Preprint*, arXiv:2511.06078.
- Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang, and Mrinmaya Sachan. 2025. [Can LLMs Effectively Simulate Human Learners? Teachers’ Insights from Tutoring LLM Students](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 100–117, Vienna, Austria. Association for Computational Linguistics.
- Kaushal Kumar Maurya, KV Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. [Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors](#). *arXiv preprint*. ArXiv:2412.09416 [cs].
- Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2023. [Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students’ text revision, motivation, and positive emotions](#). *Computers and Education: Artificial Intelligence*, 6:100199.
- Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. 2024. [Closing the loop: Learning to generate writing feedback via language model simulated student revisions](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16636–16657, Miami, Florida, USA. Association for Computational Linguistics.
- Bang Nguyen, Tingting Du, Mengxia Yu, Lawrence Angrave, and Meng Jiang. 2025. [QG-SMS: Enhancing test item analysis via student modeling and simulation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26152–26168, Vienna, Austria. Association for Computational Linguistics.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. [Large Language Models are Students at Various Levels: Zero-shot Question Difficulty Estimation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177, Miami, Florida, USA. Association for Computational Linguistics.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative Agents: Interactive Simulacra of Human Behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23, pages 1–22, New York, NY, USA. Association for Computing Machinery.
- Henry Pit. 2025. [Henry at BEA 2025 shared task: Improving AI tutor’s guidance evaluation through context-aware distillation](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1164–1172, Vienna, Austria. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentencebert: Sentence embeddings using siamese bert-networks](#). *Preprint*, arXiv:1908.10084.
- Debdeep Sanyal, Agniva Maiti, Umakanta Maharana, Dhruv Kumar, Ankur Mali, C. Lee Giles, and Murari Mandal. 2025. [Investigating pedagogical teacher and student LLM agents: Genetic adaptation meets retrieval-augmented generation across learning styles](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13348–13389, Suzhou, China. Association for Computational Linguistics.
- Mayank Sharma and Jason Zhang. 2025. [Decoding Actionability: A Computational Analysis of Teacher Observation Feedback](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 898–907, Vienna, Austria. Association for Computational Linguistics.
- Sergi Solera-Monforte, Pablo Arnau-González, and Miguel Arevalillo-Herráez. 2024. [Be my mate: Simulating virtual students for collaboration using large language models](#). In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 1–3, Tokyo, Japan. Association for Computational Linguistics.
- Andreas Säuberli, Diego Frassinelli, and Barbara Plank. 2025. [Do LLMs Give Psychometrically Plausible Responses in Educational Assessments?](#) In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA*

- 2025), pages 266–278, Vienna, Austria. Association for Computational Linguistics.
- Anaïs Tack and Chris Piech. 2022. [The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues](#). *arXiv preprint*. ArXiv:2205.07540 [cs].
- Xiao Tan, Gary Cheng, and Man Ho Ling. 2025. [Artificial intelligence in teaching and teacher professional development: A systematic review](#). *Computers and Education: Artificial Intelligence*, 8:100355.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Ryuichi Uehara. 2024. [Enhancing Role-Playing Capabilities in Persona Dialogue Systems through Corpus Construction and Evaluation Methods](#). In *Proceedings of the 20th Workshop of Young Researchers’ Roundtable on Spoken Dialogue Systems*, pages 73–75, Kyoto, Japan. Association for Computational Linguistics.
- Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024a. [Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.
- Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024b. [Artificial intelligence in education: A systematic literature review](#). *Expert Systems with Applications*, 252:124167.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. [Simulating classroom education with LLM-empowered agents](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.
- Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2025. [Can Model Uncertainty Function as a Proxy for Multiple-Choice Question Item Difficulty?](#) In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11304–11316, Abu Dhabi, UAE. Association for Computational Linguistics.

A Example Conversation Thread

We provide an example of a conversation thread used for our evaluation and experiments. The LLM Annotations of each category (i.e., Yes, To Some Extent, No) are provided to showcase the difference in quality of a tutor’s response.

Conversation History

Tutor: Hi, could you please provide a step-by-step solution for the question below?

The question is: *Elliott is trying to walk 10,000 steps a day. He finished half of his steps on his walks to and from school and did another 1,000 steps going for a short walk with his friend. He also went for a short jog around the block and realized that after he had finished his jog, he only had 2,000 steps left to take. How many steps did Elliott take during his jog?*

Student: Elliott finished half of his steps on his walks to and from school, so he took $10,000/2 = 5,000$ steps.

Adding the 1,000 steps he took with his friend, he has taken $5,000 + 1,000 = 6,000$ steps. Subtracting 6,000 from his goal of 10,000, he has $10,000 - 6,000 = 4,000$ steps left to take. Therefore, he took 4,000 steps during his jog.

Tutor: Can you tell me how you got to your answer?

Student: Sure. I started by figuring out how many steps Elliott had taken before his jog. I knew he had finished half of his steps on his walks to and from school, so I calculated $10,000/2 = 5,000$ steps. Then I added the 1,000 steps he took with his friend, giving $5,000 + 1,000 = 6,000$ steps. After that, I subtracted 6,000 from 10,000 so I had $10,000 - 6,000 = 4,000$ steps left. This meant that he had taken 4,000 steps during his jog.

A.1 Ground Truth Solution

Ground Truth

Elliott took half of his steps on his school journey, which was $10,000/2 = 5,000$ steps.

Adding this to his short walk, he has taken $5,000 + 1,000 = 6,000$ steps.

He only has 2,000 steps left, which means that during the day, he has taken $10,000 - 2,000 = 8,000$ steps.

The difference between these is therefore the steps he took during his jog: $8,000 - 6,000 = 2,000$ steps.

A.2 LLM Annotations

- **Phi3:** Great job! To further enhance your understanding, let's explore more examples and practice solving similar problems together.

- Providing Guidance: No
- Actionability: No

- **Expert:** But after his jog he still had 2,000 steps left.

- Providing Guidance: Yes
- Actionability: Yes

- **Sonnet:** Great job showing your work, but let's take another look at the last part of the problem statement to make sure we're considering all the information given.

- Providing Guidance: To some extent
- Actionability: To some extent

B Variants of Prompts Used

Below is the main prompt used to obtain the experimental results reported in this study. Here, we instruct the model to respond how it would naturally do under a student persona, without constraints such as "last tutor response".

Model Behaviour Prompt - Main

You are a novice **student** asking your tutor for help on different problems. Your goal is to respond naturally to the tutor's last message after you understand the context of the entire conversation history. You must base your response **on your current understanding of the problem and the dialogue so far**. Instructions: (1) Read the context carefully. (2) Respond as a student would, showing your **thought process or understanding**. (3) Your response should emulate what a real student might say next. (4) Keep your response concise but realistic. (5) Ensure your response is relevant to the topic. **Now, respond as a student.**

To evaluate the stability of the results across prompts, we conducted multiple runs using different instructions as student persona prompts. We then introduced the notion of "last tutor response" as anchoring. In addition to having the model respond directly as a student, we also experimented with having it predict the answer after receiving guidance from the tutor. Below, we present some of the alternative prompts that were tested. First, we paraphrased the main prompt to control for the tested models' sensitivity to specific tokens. Next, to better represent student's thinking, we specifically induce this via the Thinking prompt. Finally, instead of responding as a novice student, we specifically guide the model to produce the correct solution based on the available hints provided in the conversation.

Model Behaviour Prompt - Paraphrase

You are playing the role of a novice student asking questions to your tutor. Respond to the latest tutor message by considering the Last Tutor Response and what you currently understand. Instructions: (1) Understand the tutor's last message thoroughly. (2) Reply like a student would, reflecting your thoughts or reasoning. (3) Keep your response brief, realistic, and on-topic. Respond now as a student.

Model Behaviour Prompt - Thinking

Imagine you are a student working through a challenging concept with your tutor. Your task is to think out loud, speculate, or propose ideas, showing how you process the information. Use the tutor's last message as a reference and draw on your current understanding. Responses should feel authentic, reflective, and relevant. Respond now as the student.

Model Behaviour Prompt - Answer Prediction

You are a student attempting to solve a problem after receiving guidance from your tutor. Your goal is to predict the answer or next step based on the tutor's last message and your current understanding of the topic. Think carefully about the hints or explanations provided, and respond as a student would when reasoning toward a solution. Keep your reply realistic, concise, and relevant. Now respond as the student, predicting the answer.

No meaningful difference was observed across all runs.