EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework

Anonymous ACL submission

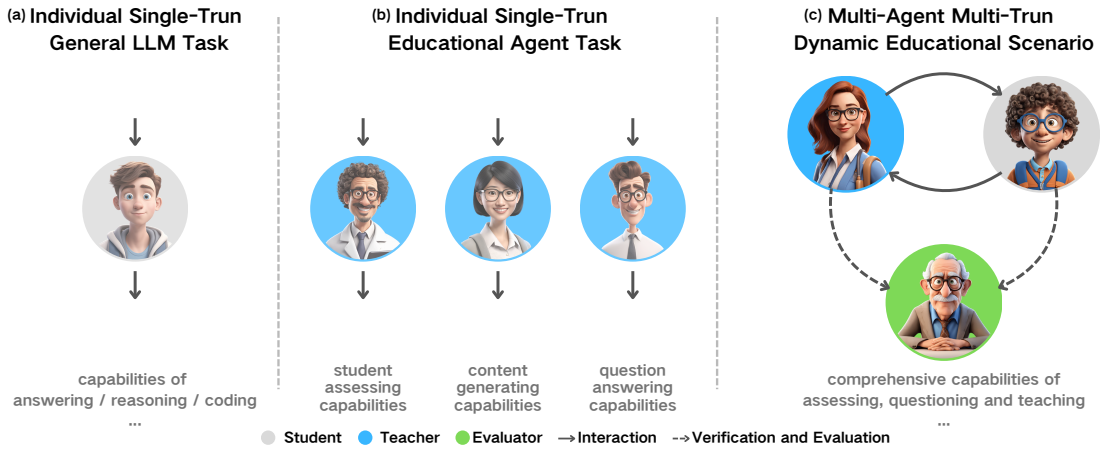

Figure 1: The evolution of LLMs in education: from individual single-turn tasks to dynamic educational scenarios simulating authentic teaching interactions. Three stages (left to right) depict the shift from isolated capabilities: (a) (b) to comprehensive teaching capabilities, (c) enabled by the EducationQ multi-agent dialogue framework.

Abstract

While Large Language Models (LLMs) demonstrate significant capabilities across domains, existing benchmarks focus primarily on knowledge and reasoning abilities, leaving a critical gap in evaluating their teaching capabilities—particularly in managing real-time instructional interactions and adapting pedagogical strategies to student needs. This paper introduces EducationQ, a novel multi-agent dialogue framework that systematically evaluates LLMs' teaching capabilities through dynamic informal formative assessment (IFA) scenarios. The framework employs a triadic interaction model comprising specialized teacher, student, and evaluator agents to capture the nuanced dynamics of educational exchanges. Using a curated dataset of 1,498 questions spanning multiple disciplines and difficulty levels, we evaluated 14 state-of-the-art LLMs. The findings challenge conventional assumptions that larger models or general capabilities inherently lead to superior teaching performance. Notably on GPQA Diamond, Teacher Llama 3.1 70B Instruct achieved significant student learning gains (12.63% improvement) through sophisticated questioning strategies, and Teacher Gemini 1.5 Pro 002 demonstrated robust performance (7.58% improvement) through adaptive feedback mechanisms—underscoring the importance of targeted teaching approaches. Quantitative metrics and qualitative dialogue analyses reveal that successful LLMs-as-teachers prioritize focused strategies and adaptive interactions aligned with established educational theories rather than broader knowledge repositories. The work contributes both a systematic framework for evaluating AI teaching capabilities and empirical insights for developing effective educational applications, bridging the gap between AI capabilities and educational needs[1].

1 Introduction

Large Language Models (LLMs) are revolutionizing various domains, sparking significant interest in their potential to transform education through personalized learning and automated feedback (Memarian and Doleck, 2023). The evolution of LLMs in educational applications has progressed from simple question-answering to increasingly sophisticated teaching capabilities (Figure 1). While recent research has explored their applications in specific teaching tasks—including question generation (Olney, 2023; Shridhar et al., 2022), automated assessment (Nye et al., 2023; Patil et al., 2024), feedback provision (Cohn et al., 2024), and teach-

[1]Code and dataset will be released once acceptance.

ing support through natural dialogue (Zha et al., 2024; Liu et al., 2024)—current benchmarks predominantly assess isolated capabilities like knowledge acquisition, reasoning, and task completion. This narrow focus fails to evaluate core teaching functions essential for effective education: guiding learning processes, facilitating knowledge construction, organizing educational activities, providing personalized feedback, and scaffolding skill development (Palincsar, 1998; Hmelo-Silver and Barrows, 2006; Mercer and Littleton, 2007; Wood et al., 1976).

Existing LLM evaluation approaches - whether through closed-ended questions, open-ended responses, or multi-turn dialogues - present fundamental limitations in assessing teaching capabilities. Current benchmarks predominantly rely on closed-ended assessments, which enable efficient automation but fail to capture the complexity and teacher agency in educational interactions. While open-ended evaluation could better reflect teaching dynamics, it faces significant challenges in scalability and consistency due to reliance on human judgment. Multi-turn dialogue frameworks, despite better capturing interactive complexity, lack specific mechanisms for evaluating teaching effectiveness. These limitations particularly impact teaching evaluation: benchmarks neither capture teachers' active role in questioning, assessment, and real-time adaptation, nor provide scalable solutions for assessing teaching quality.

To address these challenges, we propose EducationQ, a novel multi-agent dialogue framework that incorporates formative assessment into the evaluation of LLMs' teaching capabilities. Formative assessment—a continuous process of evaluating learner progress, identifying gaps, and adjusting teaching strategies (Wiliam, 2011)—is essential for personalized instruction. It bridges the gap between current abilities and potential, enhances learning outcomes, and promotes educational equity through AI (Pardo et al., 2019; Ruiz-Primo and Furtak, 2007; U.S. Department of Education, Office of Educational Technology, 2023; Allal and Pelgrims Ducrey, 2000). In classroom settings, informal formative assessments (IFAs) are a common practice during instructional dialogues, where teachers pose questions, assess student understanding, and provide timely feedback and guidance (Sezen-Barrie and Kelly, 2017; Guskey, 2005).

The EducationQ framework models these interactions through a triadic system of teacher, student, and evaluator agents, simulating cyclical teacher-student interaction. This design captures teachers' agency in employing diverse strategies and navigating complex educational contexts while enabling automated evaluation of dialogue quality. To support this framework, we curated a robust dataset of 1,498 questions from established benchmarks GPQA and MMLU-Pro, spanning diverse disciplines and difficulty levels. We employed a mixed-methods approach to comprehensively evaluate LLMs' teaching capabilities. Our framework evaluates teaching effectiveness through structured interactions, quantifying teachers' capabilities from an outcome-aligned perspective (Gitomer and Duschl, 2007) and analyzing pedagogical strategies with the evaluator agent.

Our analysis yielded several key findings. Quantitatively, we observed that superior performance in general knowledge benchmarks does not predict teaching effectiveness, with some smaller open-source models outperforming larger commercial ones. And qualitative analysis of teaching dialogues highlights distinct pedagogical strategies contributing to these outcomes.

Our findings reveal model-specific teaching strengths. Llama 3.1 70B Instruct achieved balanced and superior teaching performance through sophisticated questioning strategies, achieving 11.01% improvement across all evaluation questions and up to 24% in individual subjects. Gemini 1.5 Pro 002 achieved 7.48% improvement by providing targeted instructional feedback. OpenAI o1-mini excelled in reasoning-intensive subjects, while Llama 3.1 70B Instruct dominated knowledge-intensive disciplines.

This work advances the field of AI in education through the major contributions:

- A theoretical framework integrating formative assessment and Vygotsky's (1978) learning theory to evaluate educational LLMs.
- A multi-agent dialogue methodology for simulating and assessing authentic teaching interactions.
- A high-quality educational dataset comprising standardized tests and re-annotated teacher-student dialogues with pre/post-test results (14,980 five-round interactions).
- Vast empirical evaluations demonstrating significant student learning gains (up to 12.63% improvement on the GPQA Diamond test set).

2 Related Work

2.1 LLM Evaluation

Task-oriented performance benchmarks like MMLU (Hendrycks et al., 2021b), MMLU-Pro (Wang et al., 2024b), and GPQA (Rein et al., 2023) employ closed-ended questions to evaluate domain knowledge and reasoning abilities. Similarly, MATH (Hendrycks et al., 2021a) examines mathematical reasoning, while HumanEval (Chen et al., 2021) tests programming capabilities.

Instruction following benchmarks such as IFEval (Zhou et al., 2023), FLAN (Wei et al., 2022), Self-Instruct (Wang et al., 2022), and NaturalInstructions (Wang et al., 2022) assess LLMs' ability to comprehend and execute directives through open-ended responses.

Human preference alignment benchmarks like MT-Bench and Chatbot Arena (Zheng et al., 2023) evaluate interaction quality through human judgment, they prioritize general user satisfaction over educational outcomes.

2.2 LLM-Enhanced Benchmark Development

Recent research has increasingly incorporated LLMs as agents in benchmark datasets, tasks, and analysis. For instance, MMLU-Pro employs GPT-4-Turbo to expand distractor options, enhancing test stability (Wang et al., 2024b). Benchmarks Self-Evolving (Wang et al., 2024a) utilizes LLMs to extend existing benchmark sets, reducing data contamination while increasing stability and granularity. Dr.Academy (Chen et al., 2024) leverages GPT-4 to evaluate generated content's consistency, relevance, coverage, and representativeness.

LLMs' human-like behavior has led to their use in simulating human judgment, test-taking, and feedback provision. Zheng et al. (2023) demonstrated how human-aligned GPT-4 could replace human judges in MTBench, reducing crowdsourcing costs while maintaining evaluation quality.

2.3 LLM-Based Student Modeling

Recent work has explored using LLMs to simulate student behavior and interactions. Xu & Zhang (2023) investigated the feasibility of using generative students to test educational materials. Markel et al. (2023) employed LLMs to simulate student dialogues for teacher training. Lu & Wang (2024) found that profile-based generative students closely mirror human student performance in MCQ responses. Jin et al. (2024) proposed TeachTune, a framework generating pedagogical agent dialogues with diverse simulated student profiles for human evaluation, complementing our automated fixed-student-model assessment approach.

3 Dataset

We constructed our evaluation datasets, as summarized in Table 1, by systematically curating questions from two well-established benchmarks: GPQA (n=448), featuring domain expert-authored questions, and MMLU-Pro (n=12,032), containing reasoning-intensive questions with enhanced robustness through 10-option design across 14 educational categories. These datasets span undergraduate to PhD-level content, providing a rigorous foundation for evaluating teaching capabilities.

Data Source	Count	Extracted Dataset	Count
GPQA	448	GPQA Diamond	198
MMLU-Pro	12,032	MMLU-Pro Stratified	1,300
		Total	1,498

Table 1: Dataset construction and distribution statistics.

To optimize both assessment quality and efficiency, we focused on two carefully selected subsets: (1) GPQA Diamond (n=198), an expert-validated subset of GPQA with empirically verified difficulty (demonstrated by < 33% correct response rate among non-experts), and (2) our newly constructed MMLU-Pro Stratified (n=1,300).We developed MMLU-Pro Stratified through systematic sampling based on performance analysis of the top 10 models from published evaluation results[2] (accessed September 2024). As visualized in Figure 2, we calculated mean accuracy rates across all valid responses for each question, excluding null or malformed outputs, to assign difficulty ratings. After removing the "other" category to ensure disciplinary clarity, we stratified the remaining questions into 10 difficulty levels using 10% intervals and sampled the first 10 questions from each subject-difficulty combination.

The 1,498-question dataset attained a 47.73% baseline accuracy with Llama 3.1 70B Instruct as the student agent, providing a reference for teaching effectiveness. The distribution simulates diverse educational scenarios, with balanced representation across difficulty tiers and disciplines ensuring comprehensive analytical coverage.

[2]https://github.com/idavidrein/gpqa.

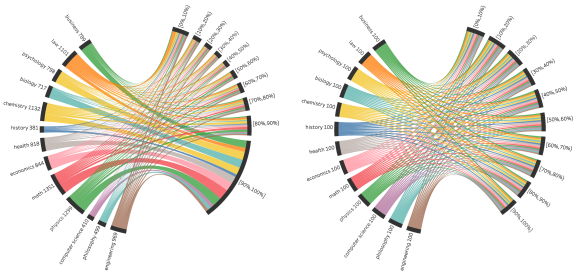

Figure 2: Dataset distribution across 13 academic disciplines and 10 difficulty levels (presented by accuracy rates of 10 high-performing LLMs). Left: original MMLU-Pro; Right: MMLU-Pro Stratified.

4 EducationQ Multi-Agent Framework

Our methodology employs three distinct agents: the teacher agent under evaluation, the student agent participating in standardized tests and IFA dialogues, and the evaluator agent providing analysis, , as illustrated in Figure 3.

4.1 Student Agent

The student agent is prompted (see Appendix A.1) to focus on specific subjects, analyze problems, and express thoughts and uncertainties, mimicking authentic student behavior. We implement soft token limits rather than hard cutoffs to maintain natural response patterns. Llama 3.1 70B Instruct (GPQA Diamond 46.97%) serves as our student agent due to its open-source availability for reproducibility, strong instruction-following capabilities (86.96 IFEval), and balanced performance-cost ratio at 70B parameters.

Ablation studies, as showed in Table 2 using Qwen 2.5 72B Instruct (IFEval 86.38; GPQA Diamond 45.45%) and Mistral Nemo 12b (IFEval 62.03; GPQA Diamond 35.35%) as alternative student models showed negligible impact on experimental rankings, suggesting our methodology effectively isolates teacher model performance differences independent of student model selection.

4.2 Teacher Agent

Teacher agents are prompted (see Appendix A.2) to conduct dynamic assessment of student thinking processes and dialogue performance, employing probing questions to gauge understanding and promote thinking, providing feedback, and offering necessary corrections(Sezen-Barrie and Kelly, 2017).

To prevent direct answer disclosure, we restrict teacher agents' access to question options and explicitly require them to guide without revealing answers.

4.3 Evaluator Agent

The evaluator agent is prompted (see Appendix A.3) as an education assessment expert well-versed in pedagogical theory and practice. It validates the dialogues and evaluates teaching dialogues according to specified dimensions and compares teacher performances to determine superior approaches.

The assessment framework comprises 17 distinct scoring dimensions, including teacher-focused metrics (questioning, assessment, feedback) and

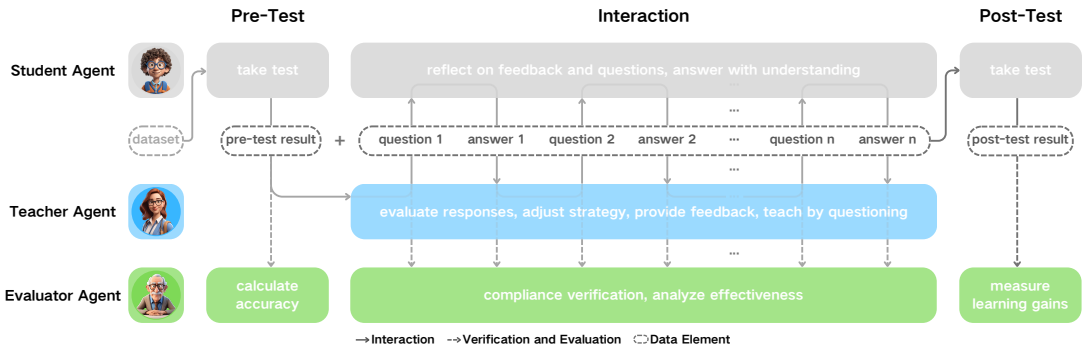

Figure 3: The formative assessment interaction flow in the EducationQ framework, detailing the multi-agent multi-turn dialogue implementation shown in Figure 1(c).

Teacher / Student	Llama 3.1 70B Instruct						Qwen 2.5 72B Instruct						Mistral Nemo					
	Accuracy (%)			Metrics			Accuracy (%)			Metrics			Accuracy (%)			Metrics		
	Pre	Post	Δ	PNIR	CSS	UIC	Pre	Post	Δ	PNIR	CSS	UIC	Pre	Post	Δ	PNIR	CSS	UIC
Llama 3.1 70B Instruct	46.97	59.60	12.63	0.26	0.26	22	46.97	55.05	8.08	0.27	0.26	8	46.97	51.52	4.55	0.47	0.24	5
Qwen 2.5 72B Instruct	45.45	54.04	8.59	0.06	0.22	13	45.45	50.00	4.55	0.18	0.24	4	45.45	47.98	2.53	0.17	0.25	2
Mistral Nemo	35.35	42.42	7.07	0.42	0.18	17	35.35	37.88	2.53	0.72	0.19	13	35.35	35.35	0.00	1.00	-	8

Dataset: GPQA Diamond, Pre: Pre-test accuracy, Post: Post-test accuracy, Δ: Absolute learning gain, PNIR: Positive-Negative Impact Ratio (lower is better), CSS: Cross-subject Stability (lower is better), UIC: Unique Improvement Count.

Table 2: Student Agent Ablation Study Based on GPQA Diamond.

student-impact measures (metacognitive reflection, knowledge dimension, etc) (Krathwohl, 2002; Looney, 2011; Wilen, 1987; Wass and Golding, 2014). Given the exploratory nature of this component, we did not address dimensional overlap.

4.4 Interaction Protocol

Our teaching interaction design simulates informal formative assessment (IFA) scenarios in classroom settings. In these contexts, the boundaries between curriculum, instruction, and assessment become fluid (Duschl and Gitomer, 1997), with teachers enhancing students' mastery of learning objectives through continuous dialogue, ultimately leading to improved summative assessment performance (Ruiz-Primo and Furtak, 2007).

In our framework, pre-test and post-test correspond to standardized summative assessments, while the multi-turn interactions represent classroom IFAs. The difference in accuracy between these assessments, termed Absolute Learning Gain (ALG) 1, reflects student performance changes before and after teacher dialogue (McGrath et al., 2015), providing a reliable measure of overall teaching effectiveness.

4.5 Pre-Test

The pre-test establishes the student agent's initial knowledge baseline while providing teachers with preliminary insights through chain-of-thought reasoning patterns. To ensure broader applicability and stability, we conducted pre-test evaluation following official MMLU-Pro and GPQA Diamond benchmark protocols and parameters.

4.6 Interaction

Dialogues proceed question by question, with teachers receiving message-format access to question content, student responses, and correctness judgments before initiating the first interaction round. Each teacher-student exchange constitutes one round, with five rounds per question.

4.7 Post-Test

To maintain compatibility with existing benchmarks, we employed MMLU-Pro and GPQA evaluation protocols rather than student agent assessment. The post-test incorporated pre-test reasoning records and subsequent teacher-student dialogue content via message format while maintaining consistent parameter settings.

4.8 Content Boundary Design

To prevent direct answer disclosure, we implemented the following constraints: (a) Teacher agents cannot access answer options, relying solely on student reasoning patterns and correctness judgments for guidance. (b) Students cannot access pre-test correctness judgments during dialogue, learning exclusively through teacher interaction. (c) Students retain access to complete question content including options, enabling learning through experience association.

4.9 Interaction Parameters and Constraints

Ablation studies identified optimal parameters of 150 tokens per dialogue round across five rounds, balancing effectiveness and efficiency. Increasing token limits to 250 yielded no significant learning gains, while reducing teacher dialogue to 70-100 tokens degraded teaching performance. Doubling rounds to 10 (with halved student token limits) increased computational costs without surpassing the effectiveness of the 5-round, 150-token configuration. In final experiments, teacher responses averaged 73.6 tokens, with student responses averaging 260 tokens.

4.10 Data Quality Verification

Two automated retry mechanisms ensure data integrity: (1) Empty Response Detection: Triggers on zero token count, indicating model output failure. (2) Anomalous Output Detection: Activates when token counts significantly exceed normal ranges (>80% of 1024 tokens for dialogue or >80% of 2048 tokens for test answers).

These mechanisms automatically retry with a five-attempt limit per question. Normal response token counts averaged: 73.6 for teacher dialogue, 260 for student responses, and 425 for test answers. All retry-triggering cases underwent manual review for root cause analysis and validation, ensuring interaction data reliability.

5 Evaluation Metrics

We developed a comprehensive evaluation framework considering both quantitative performance and teaching stability:

1. Absolute Learning Gain (ALG): Measures the direct improvement in student performance:

$$ALG = ACC_{post} - ACC_{pre} \quad (1)$$

where ACC_{post} and ACC_{pre} represent the accuracy scores in post-test and pre-test, respectively. This metric reflects the overall teaching effectiveness and enables direct comparison with conventional benchmarking methods.

2. Positive-Negative Impact Ratio (PNIR): Evaluates the consistency of teaching effectiveness:
$$PNIR = \frac{N_{neg}}{N_{pos}} \quad (2)$$
where N_{neg} and N_{pos} represent the number of negative and positive teaching impact cases, respectively. Lower PNIR indicates more stable teaching performance.

3. Cross-subject Stability (CSS): Measures the standard deviation of learning gains across subjects:
$$CSS = \sigma(SLGPD) \quad (3)$$
where σ denotes the standard deviation and $SLGPD$ represents Subject-wise Learning Gains Percentage Distribution. A lower CSS value indicates more consistent cross-subject teaching capability.

4. Unique Improvement Count (UIC): Identifies questions where only one specific teacher model achieved improvement:
$$UIC = Count(QUI) \quad (4)$$
where QUI denotes the set of Questions with Unique Improvement, representing cases where only a single teacher model demonstrated enhanced performance. This metric helps identify specialized teaching capabilities of different models.

6 Experimental Setup

We evaluated both closed and open-source LLMs across different companies and scales, selecting models with varying performance levels on MMLU and GPQA benchmarks to ensure comprehensive coverage.

All experiments were conducted through online providers, with provider selection based on documented performance metrics. Detailed specifications are provided in Appendix C.

Our experiments generated 19,474 valid dialogue sequences across 1,498 questions, along with 5,032 qualitative analyses from the evaluations of 296 dialogues across 17 educational dimensions.

7 Results

This section analyzes LLMs' teaching performance through our primary metrics and evaluator-based analysis as in Table 3. Our findings reveal that teaching ability does not correlate linearly with general reasoning capabilities or model scale, with different models exhibiting distinct pedagogical strengths and strategies.

7.1 Overall Performance

In terms of overall teaching effectiveness, Llama 3.1 70B Instruct demonstrated superior capability, achieving average learning gains of 10.9%. Gemini 1.5 Pro 002 followed closely with 7.54% improvement. These results indicate that smaller open-source models can surpass larger commercial models through effective teaching strategies. Performance showed high correlation across both datasets (r=0.871, p<0.001), demonstrating our methodology's reliability.

7.2 Subject-Specific Performance

Analysis across disciplines revealed distinct specializations among models. Llama 3.1 70B Instruct excelled in knowledge-intensive subjects, leading in psychology (ALG=18%), health sciences (ALG=24%), and law (ALG=11%). OpenAI o1-mini dominated physics (ALG=8.6%) and mathematics (ALG=9%), demonstrating strength in logical reasoning and problem-solving. Gemini 1.5 Pro 002 showed particular prowess in applied disciplines like business (ALG=8%) and economics (ALG=9%), reflecting superior integration of theoretical knowledge with practical applications. Additionally, Hermes 3 Llama 3.1 70B led in engineering (ALG=10%), while Qwen 2.5 72B Instruct topped chemistry in MMLU-Pro (ALG=11%).

In terms of cross-subject stability (CSS), Gemini 1.5 Pro 002 (CSS=0.031) and Llama 3.1 70B Instruct (CSS=0.041) demonstrated the most consistent performance across disciplines.

7.3 Performance Across Difficulty Levels

Llama 3.1 70B Instruct showed the most stable performance across difficulty levels (σ=0.032), closely followed by Gemini 1.5 Pro 002 (σ=0.043). Most LLM teachers performed best with relatively simple questions (prior accuracy $\tilde{0}.8$), with these improvements accounting for approximately 20% of total gains, suggesting strength in reinforcing well-understood concepts.

However, the Llama 3.1 series (70B and 8B models) exhibited a distinctly different pattern, achieving peak performance at medium difficulty levels (prior accuracy $\tilde{0}.5$), accounting for 27% and 19%

Table 3: Performance Comparison of Different Language Models

Model	GPQA Diamond Accuracy (%)			MMLU-Pro Stratified Accuracy (%)			Overall Accuracy (%)			Additional Metrics		
	Pre	Post	Δ	Pre	Post	Δ	Pre	Post	Δ	CSS	PNIR	UIC
Llama 3.1 70B Instruct	46.97	**59.60**	**12.63**	47.85	**58.62**	**10.77**	47.73	**58.74**	**11.01**	4.14	**0.18**	**37**
Gemini 1.5 Pro 002	46.97	54.55	7.58	47.85	55.31	7.46	47.73	55.21	7.48	**3.02**	0.40	**37**
Llama 3.1 405B Instruct	46.97	*55.05*	*8.08*	47.85	*53.69*	*5.85*	47.73	*53.87*	*6.14*	*4.54*	0.24	9
OpenAI o1-mini	46.97	56.57	9.60	47.85	53.12	5.27	47.73	53.57	5.84	5.14	*0.25*	7
Qwen 2.5 72B Instruct	46.97	*55.05*	*8.08*	47.85	52.85	5.00	47.73	53.14	5.41	5.37	0.33	7
Llama 3.1 8B Instruct	46.97	52.02	5.05	47.85	52.69	4.85	47.73	52.60	4.87	5.06	0.40	13
Hermes 3 Llama 3.1 70B	46.97	51.52	4.55	47.85	51.92	4.08	47.73	51.87	4.14	5.11	0.39	6
Mistral Nemo	46.97	51.52	4.55	47.85	51.69	3.85	47.73	51.67	3.94	5.78	0.44	*12*
Claude 3.5 Sonnet	46.97	52.53	5.56	47.85	51.38	3.54	47.73	51.54	3.81	5.86	0.30	5
WizardLM-2 8x22B	46.97	50.51	3.54	47.85	51.54	3.69	47.73	51.40	3.67	4.74	0.34	2
DeepSeek V2.5	46.97	50.51	3.54	47.85	51.08	3.23	47.73	51.00	3.27	5.14	0.46	3
Command R 08-2024	46.97	49.49	2.53	47.85	50.85	3.00	47.73	50.67	2.94	5.68	0.53	7
GPT-4o-mini	46.97	50.51	3.54	47.85	50.12	2.27	47.73	50.17	2.44	8.47	0.40	2
Phi-3.5-mini Instruct	46.97	48.99	2.02	47.85	48.92	1.08	47.73	48.93	1.20	17.23	0.69	4

Note: Pre: Pre-Test Accuracy; Post: Post-Test Accuracy; Δ: Absolute Learning Gain; CSS: Cross-subject Stability (lower is better); PNIR: Positive-Negative Impact Ratio (lower is better); UIC: Unique Improvement Count. The best results are marked in **bold**, second best results are underlined, and third best results are in *italics*.

of their respective ALGs. In contrast, their improvement rates at the 0.8 difficulty level represented only 11% of their ALGs.

This unique pattern suggests these models possess superior capability in handling challenging concepts rather than merely reinforcing easily understood material, demonstrating effectiveness in helping students breakthrough current knowledge boundaries.

7.4 Teaching Stability Analysis

Through analysis of the Positive-Negative Impact Ratio (PNIR) 2, we identified significant variations in teaching stability across models. Llama 3.1 70B Instruct demonstrated exceptional stability, generating only 36 negative cases against 200 positive improvements (PNIR = 0.18). While Gemini 1.5 Pro 002 achieved comparable positive cases (188), its higher PNIR of 75 indicated greater performance volatility. OpenAI o1-mini and Llama 3.1 405B Instruct maintained moderate stability (PNIR = 0.25 and 0.28 respectively). These findings suggest that high teaching effectiveness and stability can coexist.

7.5 Unique Improvement Analysis (UIC)

Analysis of cases where only specific teacher models achieved improvement revealed distinct pedagogical strengths. Gemini 1.5 Pro 002 and Llama 3.1 70B Instruct particularly excelled, achieving 38 and 34 unique improvements respectively, significantly outperforming other models. Their success patterns showed notable differences: Llama 3.1 70B Instruct demonstrated balanced cross-disciplinary improvements (standard deviation 0.036, peaking at 14% in psychology), while Gemini 1.5 Pro 002 showed stronger subject preferences (standard deviation 0.056, reaching 21% in biology). Notably, while OpenAI o1-mini showed modest overall performance, it achieved 3 unique improvements in engineering, suggesting specialized technical expertise.

7.6 Evaluator-Based Exploratory Analysis

To understand model performance variations, we conducted an evaluator-agent analysis of 148 unique improvement cases and their paired non-improvement controls (296 dialogues total). This investigation aimed to demonstrate the potential of LLM-based evaluators in educational dialogue assessment while exploring the distinctive teaching strategies underlying performance variations.

Initial logistic regression analysis revealed significant predictors: questioning quality for Llama 3.1 70B Instruct (Exp(B)=32.864, p=0.043) and feedback quality for Gemini 1.5 Pro 002 (Exp(B)=5227.342, p=0.019). Random forest analysis further identified Llama 3.1 70B Instruct's effectiveness correlated strongly with questioning strategies (importance score 0.35) and feedback quality (0.30), while Gemini 1.5 Pro 002's success primarily stemmed from feedback mechanisms (0.40) and adaptive teaching (0.30).

8 Case Study: Teaching Strategy Analysis

Through detailed dialogue analysis, we examined strategy differences across models in specific teaching scenarios, investigating why models with superior general capabilities might underperform in ed-

ucational interactions. Our selected cases revealed intriguing pedagogical variations.

D1: Mathematical Reasoning Examples (Question 240)

D1.1: Llama 3.1 70B Intruct

T: "Can you explain why you chose to divide..."
S: [response]
T: "... or could it be an approximation? How might this affect your calculation, and what implications would it have for determining the correct answer?"

D1.2: Llama 3.1 405B Instruct

T: "... how much will Daniel's ..."
S: [response]
T: "Can you calculate..."

D2: Legal Concept Examples (Question 961)

D2.1: Gemini 1.5 Pro 002

T: "While taking out an insurance policy could be part of a larger criminal plan, is it in itself a substantial step directly towards committing the crime of murder? Think about the specific actions needed to carry out a murder and whether purchasing insurance moves the wife closer to completing those actions."
S: [response]
T: "Good, you've identified that purchasing the insurance policy is not a substantial step. Now, let's consider the wife's other actions."

D2.2: Claude 3.5 Sonnet

T: "... In the context of attempted murder, what constitutes a 'substantial step' towards committing the crime? Can you provide an example of a substantial step that would be more clearly indicative of an attempt to commit murder than purchasing a life insurance policy?..."
S: [response]
T: "...You've provided a good example of a more substantial step. ..."

Table 4: Examples of Teacher-Student Dialogues

In analyzed dialogues, Llama 3.1 70B Instruct demonstrated sophisticated teaching techniques highly aligned with established educational theory. These cases showed carefully constructed question sequences and metacognitive support scaffolding student understanding. As illustrated in dialogue D1.1, the model employed guidance across different cognitive levels (Bloom's Taxonomy) rather than focusing solely on procedural practice. This contrasted sharply with Llama 3.1 405B Instruct's approach to the same problem (D1.2), which, despite greater general capabilities, emphasized repetitive practice over conceptual understanding. Notably, Llama 3.1 70B Instruct's progressive questioning through "can you explain why" and "how might this affect" constructed cognitive bridges between students' current understanding and target concepts, exemplifying excellent application of Zone of Proximal Development theory (Vygotsky, 1978).

Gemini 1.5 Pro 002 demonstrated strong adaptive teaching capabilities, characterized by precise diagnostic techniques and targeted, specific feedback. In dialogue D2.1, it successfully identified and addressed student misconceptions about legal concepts, using concept-definition-focused questions to prompt reconceptualization and reinforcing academic concept determination through feedback. This focused approach contrasted with Claude 3.5 Sonnet's broader methodology and formalized feedback (D2.2), which introduced multiple concepts without adequately addressing core misconceptions and provided feedback based solely on task completion. Gemini 1.5 Pro 002's rapid diagnosis of conceptual misunderstandings, immediate feedback, and timely strategy adjustments demonstrated excellent formative assessment practice.

These analyses support our quantitative findings while illustrating why larger models may underperform in teaching tasks despite broader knowledge. In observed dialogues, larger models often demonstrated broader knowledge but lacked the focused, pedagogically sound interaction strategies exhibited by Llama 3.1 70B Instruct and Gemini 1.5 Pro 002. However, these observations are based on limited case analysis, primarily intended to provide initial qualitative insights into teaching capability differences among models.

9 Conclusion

Our comprehensive evaluation of LLMs' teaching capabilities reveals two critical insights: First, smaller open-source models can outperform larger commercial models through effective pedagogical strategies, challenging conventional assumptions about model scale and teaching effectiveness. Second, successful LLM teachers excel through focused, goal-oriented interactions and adaptive teaching methods rather than broader knowledge repositories.

These findings suggest a fundamental rethinking of educational LLM development: prioritizing specialized teaching capabilities over general model scaling. The significant performance variations across models in teaching tasks indicate that traditional metrics of model capability (such as knowledge breadth or reasoning ability) poorly predict teaching effectiveness, highlighting the need for education-specific evaluation frameworks.

Limitations

Our study faces several limitations in evaluation framework, test data, and model selection. Regarding the evaluation framework, our one-on-one IFA scenario cannot fully capture the complexity of teaching roles and capabilities in practice, such as managing classroom dynamics or using student dialogue for concept explanation. Our limitation on dialogue rounds prevented comparison of different LLMs' teaching efficiency in improving ALG.

In terms of model selection, our teacher model choices did not include newer or older versions within the same series, preventing tracking of teaching capability evolution in LLM development. We also excluded multimodal models and specialized educational private models.

While our test set included advanced topics from graduate to PhD levels across multiple disciplines, we did not evaluate LLMs' teaching performance with lower-grade content, such as elementary or middle school materials.

Alignment with real-world scenarios represents another major limitation, particularly regarding student modeling and simulation fidelity. Despite basic student ablation studies, we did not employ more sophisticated generative student methods to simulate diverse age groups, cognitive levels, backgrounds, and motivations, thus not fully reflecting the complexity of real teaching situations.

Our decision not to extensively use evaluators' evaluations or incorporate qualitative indicators in primary benchmark testing meant our main metrics might not fully capture the diversity of LLMs' teaching capabilities. While evaluator-based analysis provided valuable insights, its key findings require further validation.

Model Content Limitations

During experimentation, we observed protential impacts of content policies on model evaluation. Specifically, OpenAI models (including OpenAI o1-mini and GPT-4o-mini) consistently returned NoneType responses when handling questions about the Vietnam War (Question 5048). This phenomenon, occurring only with specific content-model combinations, likely stems from provider content moderation policies.

This observation highlights a crucial limitation of commercial models in academic evaluation: content moderation policies may create gaps or biases in assessing historically or politically sensitive topics. Such constraints require careful consideration when designing educational evaluation frameworks and academic applications.

Question ID: 5048
Topic: Political Divergence During Vietnam War
Content: Description of War Impact on Society
Model Response: Consistent NoneType Returns

Table 5: Question Analysis Example

Ethics Statement

This work focuses on evaluating LLMs' teaching capabilities through automated assessment. While our framework demonstrates potential for educational applications, we acknowledge several ethical considerations:

First, our evaluation framework is designed to assess teaching capabilities rather than replace human teachers. The simulated teaching interactions should be viewed as complementary tools for understanding AI systems rather than substitutes for human-student relationships.

Second, we recognize the limitations of our single-student model approach and the potential bias in educational assessment. Our findings should be interpreted within the context of these constraints, particularly when considering real-world applications.

Our dataset is constructed from publicly available benchmarks (GPQA and MMLU-pro) following their respective terms of use and licensing agreements. We ensure proper attribution and usage of these resources in accordance with their intended research purposes.

Finally, we observed content filtering in some commercial models, highlighting the need for transparent discussion of AI systems' limitations in handling sensitive educational topics.

References

L. Allal and G. Pelgrims Ducrey. 2000. Assessment of—or in—the zone of proximal development. volume 10, pages 137–152.

M. Chen, J. Tworek, H. Jun, Q. Yuan, et al. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Y. Chen, C. Wu, S. Yan, P. Liu, H. Zhou, and Y. Xiao. 2024. Dr.academy: A benchmark for evaluating questioning capability in education for large language models. *Preprint*, arXiv:2408.10947.

C. Cohn, N. Hutchins, T. Le, and G. Biswas. 2024. A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23182–23190.

R. A. Duschl and D. H. Gitomer. 1997. Strategies and challenges to changing focus of assessment and instruction in science classrooms. *Educational Assessment*, 4:37–73.

D. H. Gitomer and R. A. Duschl. 2007. chapter 12 establishing multilevel coherence in assessment. In *Yearbook of the National Society for the Study of Education*, volume 106, pages 288–320.

T. Guskey. 2005. Formative classroom assessment and benjamin s. bloom: theory, research, and implications. Paper presented at the annual meeting, American Educational Research Association. Available at: https://files.eric.ed.gov/fulltext/ED490412.pdf.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. 2021a. Measuring mathematical problem solving with the math dataset.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

C. Hmelo-Silver and H. Barrows. 2006. Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-Based Learning*, 1(1).

H. Jin, M. Yoo, J. Park, Y. Lee, X. Wang, and J. Kim. 2024. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. *Preprint*, arXiv:2410.04078.

D. R. Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218.

J. Liu, Y. Yao, P. An, and Q. Wanliug. 2024. Peergpt: Probing the roles of llm-based peer agents as team moderators and participants in children's collaborative learning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Janet Looney. 2011. Integrating formative and summative assessment. Technical Report 58, OECD Education Working Papers.

X. Lu and X. Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pages 16–27.

J. M. Markel, S. G. Opferman, J. A. Landay, and C. Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, pages 226–236.

Cecile McGrath, Benoit Guerin, Emma Harte, Michael Frearson, and Catriona Manville. 2015. Learning gain in higher education.

B. Memarian and T. Doleck. 2023. Chatgpt in education: Methods, potentials, and limitations. *Computers in Human Behavior: Artificial Humans*, 1(2):100022.

N. Mercer and K. Littleton. 2007. *Dialogue and the Development of Children's Thinking: A Sociocultural Approach*, 1 edition. Routledge.

B. Nye, D. Mee, and M. G. Core. 2023. Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *AIED Workshops*. CEUR-WS.org.

A. Olney. 2023. Generating multiple choice questions from a textbook: Llms match human performance on most metrics. In *Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation at the AIED'23 Conference*.

A. S. Palincsar. 1998. Social constructivist perspectives on teaching and learning. *Annual Review of Psychology*, 49:345–375.

A. Pardo, J. Jovanovic, S. Dawson, D. Gašević, and N. Mirriahi. 2019. Using learning analytics to scale the provision of personalised feedback. In *British Journal of Educational Technology*, volume 50, pages 128–138.

Rajlaxmi Patil, Aditya Ashutosh Kulkarni, Ruturaj Ghatage, Sharvi Endait, Geetanjali Kale, and Raviraj Joshi. 2024. Automated assessment of multimodal answer sheets in the stem domain. *Preprint*, arXiv:2409.15749.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

M. A. Ruiz-Primo and E. M. Furtak. 2007. Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. volume 44, pages 57–84.

A. Sezen-Barrie and G. J. Kelly. 2017. From the teacher's eyes: Facilitating teachers noticings on informal formative assessments (ifas) and exploring the challenges to effective implementation. *International Journal of Science Education*, 39(2):181–212.

K. Shridhar, J. Macina, M. El-Assady, T. Sinha, M. Kapur, and M. Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149.

U.S. Department of Education, Office of Educational Technology. 2023. Artificial intelligence and future of teaching and learning: Insights and recommendations. Technical report, U.S. Department of Education, Washington, DC.

L. S. Vygotsky. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. 2024a. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *Preprint*, arXiv:2402.11443.

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N.A. Smith, D. Khashabi, and H. Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

R. Wass and C. Golding. 2014. Sharpening a tool for teaching: The zone of proximal development. *Teaching in Higher Education*, 19(6):671–684.

J. Wei, M. Bosma, V. Zhao, K. Guu, A.W. Yu, B. Lester, N. Du, A.M. Dai, and Q.V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

William W Wilen. 1987. *Questions, questioning techniques, and effective teaching*. What Research Says to the Teacher. ERIC, Washington, DC.

D. Wiliam. 2011. Formative assessment: Definitions and relationships. *Educational Assessment, Evaluation and Accountability*, 23(1):7–27.

D. Wood, J. S. Bruner, and G. Ross. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

Songlin Xu and Xinyu Zhang. 2023. Leveraging generative artificial intelligence to simulate student learning behavior. *Preprint*, arXiv:2310.19206.

S. Zha, Y. Liu, C. Zheng, J. Xu, F. Yu, J. Gong, and Y. Xu. 2024. Mentigo: An intelligent agent for mentoring students in the creative problem solving process. *Preprint*, arXiv:2409.14228.

L. Zheng, W. Chiang, Y. Sheng, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

J. Zhou, T. Lu, S. Mishra, S. Brahma, et al. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

A Agent Prompting Templates

A.1 Student Agent Prompt

The student agent uses a consistent prompting template designed to simulate authentic student learning behavior. Below we detail the exact configuration and prompting templates used in our experiments.

A.1.1 Implementation Parameters

In our experiments, we used the following configuration for the student agent (Meta-Llama-3.1-70B-Instruct):

- Maximum tokens for dialogue responses: 1,024
- Maximum tokens for test responses: 2,048
- Target token limit for standard responses: 150
- Target token limit for test responses: 1,024
- Token rerun threshold: 80%
- Maximum retries per response: 5
- Temperature: 0.0

A.1.2 Base System Message

The core system message template:

```
You are a student focusing on [CATEGORY
    ]. Analyze the question carefully,
    explain your thought process ([
    TOKEN_LIMIT] tokens or less), and
    try to apply the concepts you've
    learned to solve problems. If you're
     unsure, express your uncertainty
    and explain your reasoning.
```

A.1.3 Response Format

For test questions, responses are structured as:

```
Question: [QUESTION_TEXT]
Options:
A. [OPTION_A]
B. [OPTION_B]
C. [OPTION_C]
D. [OPTION_D]
Let's think step by step.
[REASONING_PROCESS]
The answer is (X)
```

For dialogue interactions, responses follow the format:

```
Teacher: [TEACHER_QUESTION]
Student: [STUDENT_RESPONSE]
```

The implementation details can be found in the `StudentLLM` class, specifically in the `answer_question()` and `take_test()` methods.

A.2 Teacher Agent Prompt

The teacher agent employs a structured prompting system to conduct dynamic assessment and provide guided instruction. Below we detail the configuration and prompting templates used for our teacher models.

A.2.1 Implementation Parameters

Common configuration across all teacher models:

- Maximum tokens per response: 1,024
- Target token limit for questions: 150
- Token rerun threshold: 80%
- Maximum retries per question: 5
- Temperature: 0.0

A.2.2 Base System Message

The core system message template used for all teaching interactions:

```
You are an expert teacher in [CATEGORY]
    dedicated to enhancing the student's
     understanding after analyzing the
    student's response to a pre-test.

Your task is to ask [NUM_ROUNDS] rounds
    of relevant, thought-provoking
    questions to the student. You should
     ask one new question per round (and
     if needed, provide necessary
    corrections or feedback for the
    student's previous round's answers),
     each under [TOKEN_LIMIT] tokens,
    without revealing the correct
    answers or specific details of the
    pre-test questions.

Your goal is to prepare the student for
    the post-test by fostering a deeper
    and more comprehensive understanding
     of the subject matter.
```

A.2.3 Pre-test Information Format

Pre-test results are provided in the following format:

```
Question ID: [ID]
Question: [QUESTION_TEXT]
Student's Reasoning: [REASONING]
Student's Answer: [ANSWER]
Student's Answer is Correct or Not: [
    EVALUATION]
```

A.2.4 Interaction Format

Each round of teacher-student interaction follows:

```
Teacher: [PREVIOUS_QUESTION]
Student: [STUDENT_RESPONSE]
Teacher: Generate the round [N] question
     ([TOKEN_LIMIT]
tokens or less) to promote better
    understanding:
```

The implementation details can be found in the `TeacherLLM` class, specifically in the `generate_question()` method.

A.3 Evaluator Agent Prompt

The evaluator agent is configured as an expert in educational assessment, providing detailed analysis across multiple dimensions. Below we detail the exact evaluation framework used in our experiments.

A.3.1 Implementation Parameters

Configuration for the evaluator agent (GPT-4-0806):

- Maximum tokens: 4,096
- Temperature: 0.0
- Response format: Structured JSON schema

A.3.2 Evaluation Dimensions

The evaluator assesses teaching effectiveness across three major categories:

Interaction Analysis Dimensions:

- Assessment Effectiveness
- Questioning Effectiveness
- Feedback Effectiveness
- Instructional Adaptation Effectiveness
- Learning Objective Achievement Effectiveness

Teacher Questions Analysis Dimensions:

- Question Relevance
- Cognitive Level
- Knowledge Dimension
- Question Diversity
- Scaffolding Progression
- Metacognitive Promotion

Student Responses Analysis Dimensions:

- Response Relevance
- Cognitive Level Demonstration
- Knowledge Dimension Integration
- Response Diversity
- Elaboration Progression
- Metacognitive Reflection

A.3.3 Evaluation Format

For each dimension, the evaluator provides:

```
{
   "analysis": "Detailed step-by-step analysis",
    "score": Numerical score between 1-10
}
```

A.3.4 Comparative Analysis Format

When comparing two teachers:

```
{
    "teacher_a": {
        "dimension_1": {
            "analysis": "...",
            "score": N
        }
        // ... other dimensions
    },
    "teacher_b": {
        // Similar structure
    },
    "verdict": {
      "analysis": "Comparative analysis",
        "choice": "A"/"B"/"Tie"
    }
}
```

The implementation details can be found in the `EvaluatorLLM` class, including `over_interaction_analysis()`, `teacher_questions_analysis()`, and `student_responses_analysis()` methods.

B Implementation Details

B.1 Dataset Processing

For each dataset type (MMLU-Pro, GPQA), questions are processed into a standardized format:

```
{
   "question_id": str,
   "question": str,
   "options": List[str],
   "answer": str,
   "answer_index": int,
   "cot_content": str,
   "category": str
}
```

B.2 Quality Control Mechanisms

B.2.1 Response Validation

Automatic retry mechanisms are implemented with the following criteria:

- Empty Response Detection: Zero token count
- Length Validation: >80% of maximum tokens
- Maximum Retries: 5 attempts per question
- Token Limits:
  - Teacher questions: 150 tokens
  - Student answers: 150 tokens
  - Test responses: 1,024 tokens

B.3 Scoring Guidelines

All evaluator scoring follows a 1-10 scale where:

- 1-2: Significantly below expectations
- 3-4: Below expectations
- 5-6: Meets basic expectations
- 7-8: Exceeds expectations
- 9-10: Significantly exceeds expectations

B.4 Error Handling

B.4.1 API Error Recovery

Implements exponential backoff with:

- Initial delay: 10 seconds
- Maximum delay: 320 seconds
- Maximum retries: 5

B.4.2 Response Validation

For each response:

- Format validation against expected schema
- Token count verification
- Content completeness check
- Automatic retry for invalid responses

B.5 Parallel Processing

Task parallelization implemented with:

- Maximum concurrent tasks: 5
- ThreadPoolExecutor management
- Progress tracking per teacher-student pair
- Automatic result aggregation

All implementation code, configuration files, and evaluation scripts will be made available upon acceptance.

C Model Specifications

Model	Org.	Provider	Type	Context	Params
Llama 3.1 70B Instruct	Meta	hyperbolic	bf16	32K	70B
Gemini 1.5 Pro 002	Google	Google Vertex	-	4M	-
Llama 3.1 405B Instruct	Meta	hyperbolic	bf16	8K	405B
OpenAI o1-mini	OpenAI	OpenAI	-	128K	-
Qwen 2.5 72B Instruct	Alibaba	hyperbolic	bf16	32K	72B
Llama 3.1 8B Instruct	Meta	hyperbolic	bf16	32K	8B
Hermes 3 Llama 3.1 70B	Nous	hyperbolic	bf16	12K	70B
Mistral Nemo	Mistral	DeepInfra	bf16	128K	12B
Claude 3.5 Sonnet	Anthropic	Anthropic	-	200K	-
WizardLM-2 8x22B	Microsoft	DeepInfra	bf16	66K	176B
DeepSeek V2.5	DeepSeek	deepseek	fp8	128K	-
Command R 08-2024	Cohere	Cohere	-	128K	-
GPT-4o-mini	OpenAI	OpenAI	-	128K	-
Phi-3.5-mini Instruct	Microsoft	Azure	-	128K	3.8B

Note: "-" indicates unspecified information. Context window sizes are in tokens. Org.: Organization (model developer), Provider: serving platform.

Table 6: Specifications of Language Models