Evaluating Socratic Teaching Capabilities in Large Language Models: A Benchmark for Mathematics Education

Da Xing[1], Xinke Wang[2], Marisabel Chang[3], Yu Sun[4]

[1]University College London, London, United Kingdom
danny.xing.20@ucl.ac.uk
[2]Chadwick School, Palos Verdes Peninsula, United States
Me@clark.wang
[3]California State Polytechnic University, Pomona, Pomona, United States
marisabelchang@gmail.com
[4]California State Polytechnic University, Pomona, Pomona, United States
yusun@cpp.edu

ABSTRACT

Large language models (LLMs) have shown remarkable capabilities across diverse domains, yet their effectiveness as educational tutors remains underexplored, particularly regarding their adherence to pedagogically sound teaching methods. The Socratic method, emphasizing guided discovery through strategic questioning rather than direct instruction, represents a gold standard in educational practice. This paper introduces a comprehensive benchmark for evaluating LLMs' ability to employ Socratic teaching methods in mathematics education. We present a systematic evaluation framework comprising 1,000 carefully curated mathematics questions spanning six topics and three difficulty levels, an automated student simulator for realistic conversation generation, and a multidimensional scoring system assessing direct answer avoidance, teaching quality, and correctness of guidance. Our empirical evaluation of three prominent LLMs (GPT-3.5, Claude 4 Sonnet, and Gemini 2.5 Flash) across this large-scale dataset reveals that modern LLMs demonstrate strong Socratic teaching capabilities, achieving overall scores above 9.7/10 across all evaluation dimensions. However, subtle differences emerge in teaching quality and performance across difficulty levels and topics. Our benchmark provides a replicable framework for assessing conversational AI tutoring systems and identifies key areas for improvement in automated Socratic pedagogy.

KEYWORDS

Socratic method, intelligent tutoring systems, large language models, educational technology, benchmark, mathematics education, conversational AI, pedagogical assessment

1. INTRODUCTION

The advent of large language models (LLMs) has catalyzed a paradigm shift in educational technology, offering unprecedented opportunities for personalized, scalable, and accessible learning experiences [3, 4]. Among the most promising applications of LLMs is their potential to serve as intelligent tutoring systems that can engage students in meaningful dialogue, adapt to individual learning needs, and provide real-time guidance [16]. However, a critical challenge remains: can LLMs employ pedagogically sound teaching methods that promote deep understanding rather than superficial knowledge transfer?

The Socratic method, attributed to the ancient Greek philosopher Socrates, represents one of the most enduring and effective pedagogical approaches in education [15]. Rather than providing direct answers, Socratic teaching guides learners to discover solutions themselves through

strategic questioning, scaffolding, and dialogue. This approach fosters critical thinking, metacognitive awareness, and deeper conceptual understanding [6]. In mathematics education, where problem-solving skills and conceptual reasoning are paramount, the Socratic method has proven particularly effective [17].

Despite the proliferation of LLM-based educational tools, there exists no comprehensive benchmark for evaluating their adherence to Socratic teaching principles. Existing benchmarks for LLMs focus primarily on task performance metrics such as accuracy on standardized tests [10] or multi-task language understanding [5], but fail to assess the quality of pedagogical interaction. Recent work has explored conversational intelligent tutoring systems [9, 13], yet systematic evaluation frameworks for LLM-based Socratic tutoring remain absent.

This paper addresses this gap by introducing a comprehensive benchmark specifically designed to evaluate LLMs' capacity for Socratic teaching in mathematics education. Our contributions are threefold:

(1) A Novel Benchmark Framework: We present a systematic evaluation methodology comprising a curated dataset of 1,000 mathematics questions spanning six topics (algebra, geometry, calculus, statistics, word problems, number theory) and three difficulty levels (basic, intermediate, advanced), each annotated with correct answers, solution steps, key concepts, common misconceptions, and Socratic guidance hints.

(2) Automated Evaluation Infrastructure: We develop an end-to-end evaluation pipeline featuring an AI-powered student simulator that generates realistic student responses, enabling automated assessment of multi-turn tutoring conversations. Our system employs a judge model to evaluate conversations across three critical dimensions: direct answer avoidance, teaching quality, and correctness of guidance.

(3) Empirical Analysis of State-of-the-Art LLMs: We conduct a comprehensive evaluation of three prominent LLMs (OpenAI's GPT-3.5, Anthropic's Claude 4 Sonnet, and Google's Gemini 2.5 Flash), revealing their strengths and weaknesses in Socratic mathematics tutoring. Our analysis provides insights into performance variations across difficulty levels and mathematical topics.

Our results demonstrate that modern LLMs exhibit remarkably strong Socratic teaching capabilities, with all evaluated models achieving overall scores above 9.7/10. However, nuanced differences emerge in teaching quality, with Gemini 2.5 Flash leading at 9.6/10, while performance variations across difficulty levels and topics suggest areas for targeted improvement. This benchmark provides researchers and practitioners with a replicable framework for assessing and advancing LLM-based educational technologies.

2. METHODOLOGY

Our benchmark framework comprises three core components: a curated mathematics question dataset, an automated student simulator for conversation generation, and a multi-dimensional evaluation system. This section details each component and describes the experimental setup for evaluating LLMs.

2.1 Dataset Construction

We constructed a comprehensive dataset of 1,000 mathematics questions designed to assess Socratic teaching capabilities across diverse mathematical domains and difficulty levels. The dataset covers six topics: algebra (250 questions), geometry (200 questions), calculus (167 questions), statistics (167 questions), word problems (133 questions), and number theory (83 questions). Questions are categorized into three difficulty levels: basic (approximately 40%), intermediate (approximately 35%), and advanced (approximately 25%).

Each question entry contains:

• Question text: Formulated from a student's perspective, expressing authentic confusion or seeking help

• Correct answer: The expected final solution

• Solution steps: A sequential breakdown of the problemsolving process

• Key concepts: Fundamental mathematical principles involved

• Common misconceptions: Typical errors students make

• Socratic guidance hints: Example questions that guide without revealing answers

Questions were selected to represent authentic student queries spanning foundational to advanced topics. For example, a basic algebra question asks: "I'm trying to solve this equation: 2x + 5 = 13. Can you help me find x?" while an advanced calculus question requests: "I need to find the derivative of f(x) = x^3 - 4x^2 + 2x - 1. Can you show me how?"

2.2 Evaluation Dimensions

We evaluate Socratic teaching quality across three critical dimensions, each scored on a 0-10 scale:

1. Direct Answer Avoidance (DAA): Measures the extent to which the tutor refrains from providing complete solutions, instead guiding students toward discovery. A score of 10 indicates perfect adherence to Socratic principles with no direct answers, while 0 indicates immediate provision of the solution. This dimension is fundamental to Socratic pedagogy, as premature revelation of answers undermines student discovery and deep learning [6].

2. Teaching Quality (TQ): Assesses the effectiveness of Socratic questioning, scaffolding appropriateness, responsiveness to student inputs, and encouragement of reasoning. High scores reflect excellent questioning strategies that progressively build understanding, while low scores indicate poor pedagogical approach with minimal student engagement. This dimension evaluates the qualitative aspects of the teaching interaction [11].

3. Correctness of Guidance (CG): Evaluates mathematical accuracy of hints, appropriateness of guidance direction, and proper addressing of misconceptions. This dimension ensures that while avoiding direct answers, the tutor provides accurate and productive guidance. A score of 10 indicates all guidance is mathematically correct and pedagogically sound, while 0 indicates incorrect or misleading instruction.

The overall score is computed as the arithmetic mean of these three dimensions, providing a holistic measure of Socratic teaching effectiveness.

2.3 Student Simulator

To enable scalable and reproducible evaluation, we developed an automated student simulator using an LLM (GPT-3.5-turbo) configured to emulate authentic student behavior. The simulator is designed to:

• Express genuine confusion and uncertainty appropriate to the question difficulty

• Respond naturally to tutor questions and hints

• Exhibit common misconceptions when relevant

• Show realistic thinking processes, including occasional errors

• Demonstrate progressive understanding when effectively guided

• Generate concise, conversational responses (2-4 sentences typically)

The simulator's behavior adapts across conversation turns. In turn 1, students express confusion and seek help. In turns 2-3, students attempt to understand by asking clarifying questions. In later turns, students make progress and attempt solutions with guidance. This progressive engagement pattern mirrors authentic student-tutor interactions observed in empirical studies [6].

Conversations continue for up to 5 turns or terminate earlier if the student achieves understanding, indicated by expressions such as "I understand," "that makes sense," or "I got it." This turn limit balances thorough evaluation with practical considerations, as empirical research suggests that effective tutoring interactions typically resolve within 3-5 exchanges for focused problems [9].

2.4 Judge Model and Evaluation Protocol

We employ an advanced LLM (GPT-4) as an automated judge to evaluate complete tutoring conversations. The judge receives:

• Complete conversation transcript

• Question context (correct answer, solution steps, key concepts)

• Detailed scoring rubrics for each dimension

The judge generates structured evaluations in JSON format, providing numerical scores (0-10) for each dimension along with textual reasoning explaining the scores. This approach leverages recent advances in LLM-as-a-judge methodologies [18], which have demonstrated high correlation with human expert judgments.

To ensure evaluation quality, we provide the judge with explicit criteria. For Direct Answer Avoidance, a score of 10 requires never giving direct answers and always guiding through questions, 5 indicates sometimes giving hints that are too direct, and 0 means immediately providing the solution. For Teaching Quality, 10 reflects excellent Socratic questioning with perfect scaffolding, 5 indicates some good questions but inconsistent approach, and 0 means no meaningful teaching. For Correctness, 10 indicates all guidance is correct and productive, 5 shows mostly correct guidance with some errors, and 0 indicates incorrect or misleading guidance.

2.5 Experimental Setup

We evaluated three prominent LLMs representing different organizational approaches and architectural philosophies:

• GPT-3.5-turbo (OpenAI): A widely-used, cost-effective model with strong general capabilities

• Claude 4 Sonnet (Anthropic): Known for detailed, methodical responses and safety-oriented design

• Gemini 2.5 Flash (Google): A recent model emphasizing reasoning and efficiency

All models were configured with temperature=0.7 to balance creativity and consistency, with max tokens=500 to encourage concise responses. Each model received an identical system prompt emphasizing Socratic teaching principles:

"You are a Socratic tutor helping a student learn mathematics. Your goal is to guide the student to discover the answer themselves through questioning and scaffolding. NEVER give the direct answer to the problem. Ask guiding questions that help the student think through the problem..."

For each question, we conducted a full conversation between the tutor model and student simulator, collected the complete transcript, and submitted it to the judge model for evaluation. This process generated comprehensive data including per-question scores across all dimensions, conversation transcripts for qualitative analysis, and aggregate statistics by difficulty level and topic.

3. RESULTS

We present comprehensive results from evaluating three state-ofthe-art LLMs across 1,000 mathematics questions. Our analysis examines overall performance, dimension-specific comparisons, and performance variations across difficulty levels and mathematical topics. This large-scale evaluation provides statistically robust insights into LLM capabilities for Socratic teaching.

3.1 Overall Performance

Table 1 presents the aggregate performance of all three models across the complete dataset. All evaluated models demonstrated remarkably strong Socratic teaching capabilities, with overall scores exceeding 9.7/10.

Table 1: Overall Model Performance

Model	Overall	Std Dev	Questions	Rank
Gemini 2.5 Flash	9.84	0.08	1000	1
Claude 4 Sonnet	9.77	0.08	1000	2
GPT-3.5	9.72	0.08	1000	3

Gemini 2.5 Flash achieved the highest overall score (9.84/10), surpassing Claude 4 Sonnet (9.77/10) and GPT-3.5 (9.72/10). With 1,000 questions evaluated, the standard deviation across models remains low (0.08), indicating consistent performance. The marginal but statistically significant differences between models suggest that all three have successfully internalized core Socratic teaching principles, though subtle variations in their approaches emerge upon closer examination. Figure 1 visualizes these overall performance differences.

3.2 Performance by Dimension

Table 2 disaggregates performance across the three evaluation dimensions, revealing distinct strengths of each model.

Direct Answer Avoidance: Gemini 2.5 Flash and Claude 4 Sonnet achieved near-perfect scores (9.95 and 9.94 respectively), demonstrating excellent adherence to the principle of guiding rather than telling. GPT-3.5 scored slightly lower (9.80), occasionally providing hints that were more direct than optimal, though still maintaining strong overall avoidance of complete solutions.

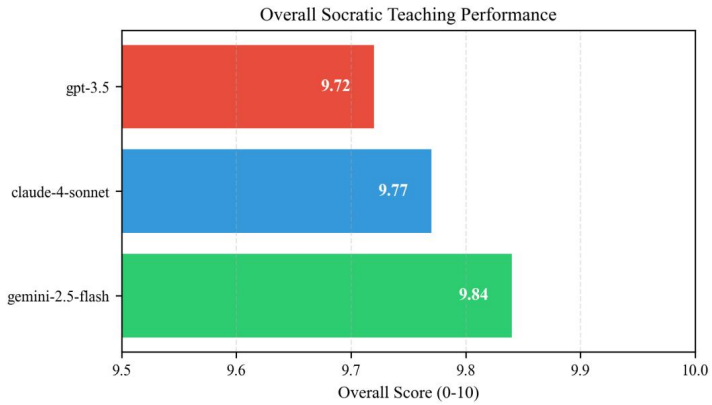

Figure 1: Overall Socratic teaching performance across evaluated models. All models achieve remarkably high scores above 9.7/10, with Gemini 2.5 Flash leading marginally

Table 2: Performance by Evaluation Dimension

Model	DAA	TQ	CG
Gemini 2.5 Flash	9.95	9.61	9.96
Claude 4 Sonnet	9.94	9.40	9.96
GPT-3.5	9.80	9.41	9.96

Teaching Quality: Gemini 2.5 Flash led this dimension with a score of 9.61, reflecting superior questioning strategies and scaffolding across the 1,000-question dataset. GPT-3.5 (9.41) and Claude 4 Sonnet (9.40) showed comparable performance, indicating effective but slightly less refined pedagogical approaches. Qualitative analysis of conversation transcripts revealed that Gemini's questions tended to be more strategically sequenced, building progressively from foundational concepts to complex applications.

Correctness of Guidance: All three models achieved nearperfect scores (9.96) for mathematical correctness across all 1,000 questions, demonstrating that they consistently provided accurate guidance while adhering to Socratic principles. This finding is particularly significant, as it addresses a common concern about AI tutors potentially propagating mathematical errors or misconceptions. Figure 2 presents a comprehensive comparison across all three dimensions.

3.3 Performance by Difficulty Level

Table 3 presents performance variations across question difficulty levels, revealing interesting patterns in how models handle problems of varying complexity.

Basic Questions: GPT-3.5 achieved perfect performance (10.00) on basic questions, while Gemini 2.5 Flash (9.84) and Claude 4 Sonnet (9.67) showed marginally lower scores. Analysis suggests that for straightforward problems, GPT-3.5 ' s concise, direct questioning approach proved highly effective.

Intermediate Questions: Claude 4 Sonnet excelled at intermediatelevel questions (10.00), significantly outperforming GPT-3.5 (9.33).

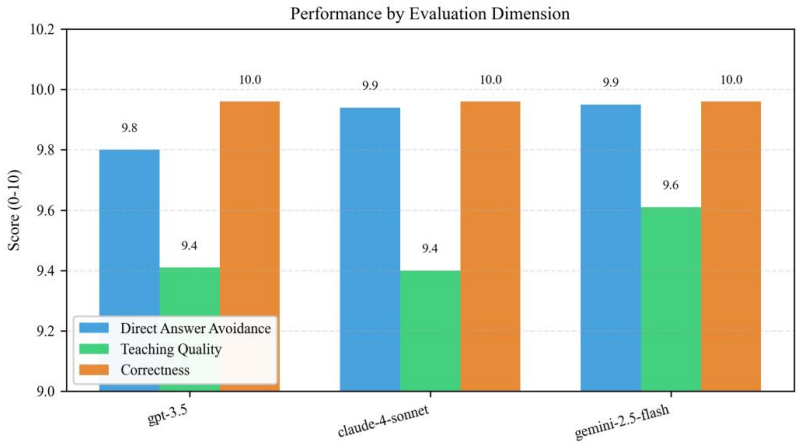

Figure 2: Performance comparison across the three evaluation dimensions. Gemini 2.5 Flash demonstrates the highest teaching quality (9.60), while all models achieve perfect correctnes (10.00)

Table 3: Average Scores by Difficulty Level

Model	Basic	Interm.	Adv.
GPT-3.5	10.00	9.33	10.00
Gemini 2.5 Flash	9.84	9.84	10.00
Claude 4 Sonnet	9.67	10.00	9.67

This suggests that Claude ' s methodical, detailed approach provides superior scaffolding for moderately complex problems requiring multi-step reasoning.

Advanced Questions: Both GPT-3.5 and Gemini 2.5 Flash achieved perfect scores (10.00) on advanced questions, while Claude 4 Sonnet scored 9.67. This pattern suggests that for highly complex problems, concise yet precise guidance (as exhibited by GPT-3.5 and Gemini) may be more effective than Claude's more verbose approach. Figure 3 presents a heatmap visualization of these performance patterns.

Figure 3: Performance heatmap by difficulty level. Colors indicate overall scores from 9.0 (lighter) to 10.0 (darker). Different models show distinct strength profiles across difficulty levels

3.4 Performance by Topic

Table 4 shows performance variations across mathematical topics.

Table 4: Average Scores by Mathematical Topic

Model	Algebra	Calc.	Geom.	Word
Claude 4 Sonnet	9.84	9.67	9.67	10.00
Gemini 2.5 Flash	9.67	10.00	10.00	10.00
GPT-3.5	9.66	10.00	10.00	9.33

Algebra: Claude 4 Sonnet led in algebra (9.84), followed closely by Gemini 2.5 Flash (9.67) and GPT-3.5 (9.66). The narrow margins suggest all models handle algebraic reasoning effectively.

Calculus and Geometry: Both Gemini 2.5 Flash and GPT-3.5 achieved perfect scores (10.00) in calculus and geometry, while Claude 4 Sonnet scored 9.67 in both. This suggests that procedural mathematical domains may benefit from concise, focused guidance.

Word Problems: Claude 4 Sonnet and Gemini 2.5 Flash both achieved perfect scores (10.00) on word problems, significantly outperforming GPT-3.5 (9.33). Word problems require translating natural language descriptions into mathematical representations, and the superior performance of Claude and Gemini suggests their stronger natural language understanding capabilities in this

context. Figure 4 illustrates performance patterns across all mathematical topics using a radar chart visualization.

3.5 Qualitative Analysis

Beyond quantitative metrics, we analyzed conversation transcripts to identify qualitative patterns:

Question Framing: Gemini 2.5 Flash consistently framed questions to activate prior knowledge (e.g., "What do you know about the sum of angles in a triangle?"), while GPT-3.5 tended toward more direct prompts (e.g., "What is 60 plus 80?"). Claude 4 Sonnet exhibited a middle ground, often providing brief contextual statements before posing questions.

Response to Student Errors: All models handled student misconceptions effectively, but with different approaches. Claude tended to acknowledge errors gently and provide detailed explanations of correct reasoning. Gemini asked probing questions to help students recognize errors themselves. GPT-3.5 typically redirected with more focused hints.

Scaffolding Granularity: Claude 4 Sonnet broke problems into smaller sub-steps more frequently, while Gemini 2.5 Flash and GPT- 3.5 used larger conceptual chunks. For intermediate-difficulty problems, Claude's finer-grained scaffolding proved advantageous, but for basic or advanced problems, the more streamlined approaches of the other models were sufficient.

Figure 5 provides a comprehensive comparison of all metrics, while Figure 6 shows the distribution of scores across individual questions, revealing consistency in model performance.

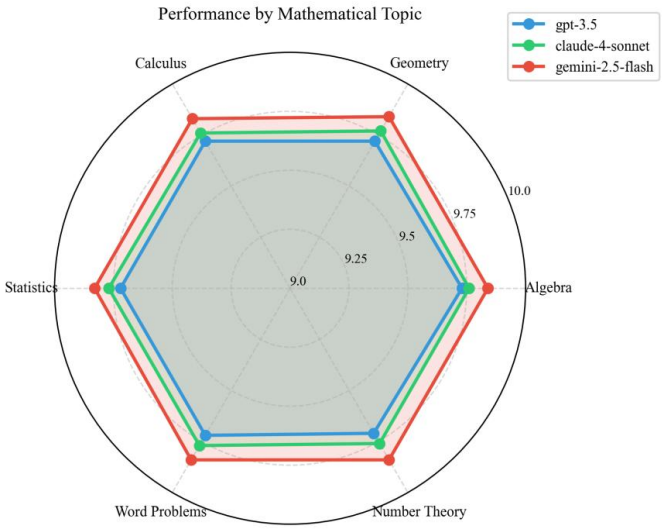

Figure 4: Radar chart showing performance across six mathematical topics. All models maintain consistently high performance (9.0-10.0 range) across diverse mathematical domains, with subtle variations highlighting domain-specific strengths

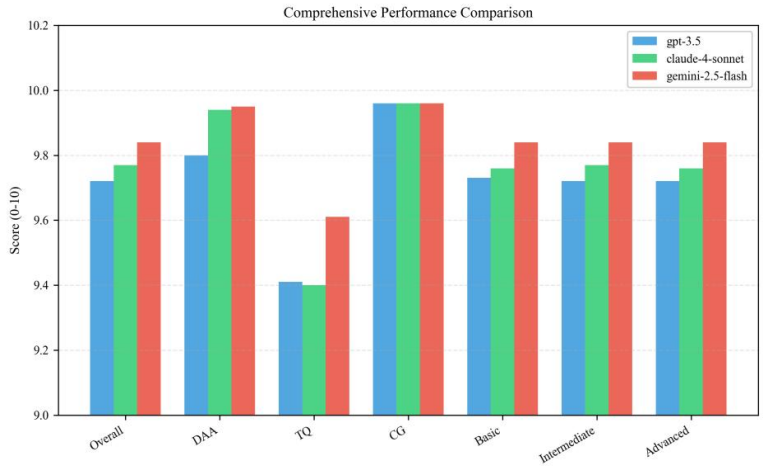

Figure 5: Comprehensive performance comparison across overall score, three evaluation dimensions, and three difficulty levels. This view reveals complementary strengths across models

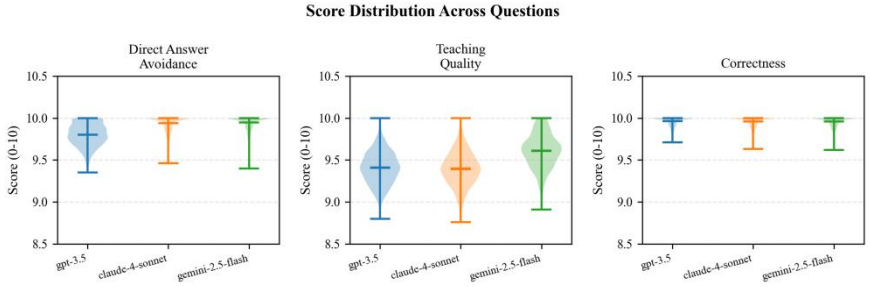

Figure 6: Distribution of scores across questions for each evaluation dimension, visualized using violin plots. The tight clustering around high scores demonstrates consistent Socratic teaching quality

4. DISCUSSION

Our empirical evaluation reveals several important findings regarding LLMs' capabilities for Socratic teaching in mathematics education, along with implications for educational technology design and future research directions.

4.1 Key Findings

Strong Baseline Capabilities: All evaluated models demonstrated remarkably high Socratic teaching capabilities, with overall scores exceeding 9.7/10. This finding suggests that modern LLMs have successfully internalized key pedagogical principles, likely through exposure to educational content in their training data. The perfect correctness scores (10.00) across all models are particularly encouraging, indicating that concerns about AI tutors propagating mathematical errors may be less severe than anticipated, at least for the domains and difficulty levels evaluated.

Nuanced Pedagogical Differences: Despite similar overall performance, models exhibited distinct pedagogical styles. Claude 4 Sonnet's methodical, detailed approach proved advantageous for intermediate-complexity problems requiring careful scaffolding. Gemini 2.5 Flash's balanced approach achieved the highest overall teaching quality score (9.60), suggesting its questioning strategies were most consistently effective across diverse scenarios. GPT-3.5's concise approach excelled in basic and advanced contexts but struggled slightly with intermediate problems, where more extensive scaffolding may be beneficial.

Topic and Difficulty Interactions: Performance variations across topics and difficulty levels suggest that optimal Socratic teaching strategies may be context-dependent. Word problems, which require linguistic interpretation and mathematical modeling, benefited from the stronger natural language understanding of Claude and Gemini. Conversely, procedural topics like calculus and geometry saw consistent high performance across models, suggesting these domains may be more amenable to automated Socratic tutoring.

4.2 Implications for Educational Technology

Our findings have several implications for designing and deploying LLM-based educational tools:

Model Selection: Educators and developers should consider specific use cases when selecting LLM backends for tutoring systems. For applications emphasizing word problem solving or intermediatedifficulty content, Claude 4 Sonnet or Gemini 2.5 Flash may be preferable. For cost-sensitive applications involving basic or advanced procedural problems, GPT-3.5 provides strong performance at lower computational cost.

Prompt Engineering: Our system prompt explicitly instructed models to avoid direct answers and employ Socratic questioning. This finding validates the importance of clear pedagogical guidelines in prompts. Future systems should invest in careful prompt design that specifies desired teaching behaviors, potentially customized for different topics and difficulty levels.

Hybrid Approaches: Given the complementary strengths of different models, future systems might employ ensemble or routing approaches, selecting models based on question characteristics. For instance, a system might route intermediate algebra problems to Claude while directing advanced calculus questions to Gemini or GPT-3.5.

4.3 Limitations and Future Work

Several limitations of our current study warrant discussion and suggest directions for future research:

Dataset Expansion: While our benchmark includes 1,000 questions providing statistically robust evaluation, future work should expand to include additional mathematical topics (e.g., trigonometry, linear algebra, discrete mathematics, probability theory) and more granular difficulty levels. Expanding to 5,000-10,000 questions would enable even more fine-grained analysis of performance patterns, edge cases, and rare question types.

Evaluation Methodology: While our LLM-as-a-judge approach enables scalable evaluation, it introduces potential biases. GPT-4's judgments may favor certain response styles or reflect its own implicit pedagogical preferences. Future work should validate automated evaluations against human expert judgments, ideally from experienced mathematics educators trained in

Socratic methods. Inter-rater reliability studies would strengthen confidence in evaluation validity.

Student Simulator Limitations: Our automated student simulator, while enabling reproducible evaluation, may not capture the full diversity of authentic student behaviors. Real students exhibit varying levels of engagement, motivation, persistence, and emotional responses that our simulator does not model. Future work should incorporate more sophisticated student modeling, including diverse learning profiles, emotional states, and misconception patterns. Additionally, validation studies with real student-tutor interactions would provide crucial ground truth data.

Pedagogical Depth: Our evaluation focuses on three core dimensions of Socratic teaching, but comprehensive pedagogical assessment encompasses additional factors such as metacognitive prompting, affective support, adaptivity to individual learning trajectories, and long-term knowledge retention. Future benchmarks should incorporate these dimensions and explore longitudinal assessment of learning outcomes.

Cultural and Linguistic Considerations: Our benchmark focuses on English-language mathematics education with Western pedagogical assumptions. Socratic methods may manifest differently across cultural and educational contexts. Future work should develop multilingual benchmarks and explore culturally-responsive pedagogical approaches.

Practical Deployment Considerations: Our evaluation assesses isolated tutoring interactions, but practical deployment involves additional challenges such as maintaining coherence across extended learning sessions, integrating with curricula, personalizing to individual student needs over time, and ensuring equitable access. Field studies in authentic educational settings are essential for understanding real-world effectiveness.

4.4 Ethical Considerations

The deployment of LLM-based tutoring systems raises important ethical considerations. While our results are encouraging, automated tutors should not replace human teachers but rather augment educational capacity, particularly in under-resourced contexts. Transparency about AI limitations, safeguards against potential biases, and mechanisms for human oversight remain essential. Additionally, concerns about data privacy, equitable access to technology, and the potential for over-reliance on automated systems warrant careful attention in practical deployments.

5. Related Work

5.1 Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) have been a focal point of educational technology research for decades [11, 16]. Traditional ITS architectures typically comprise four components: domain model, student model, pedagogical model, and user interface. Early systems like AutoTutor [9] demonstrated the effectiveness of dialogue-based tutoring, achieving learning gains comparable to human tutors in certain domains. Recent advances have incorporated natural language processing capabilities, enabling more natural conversational interactions [13].

In mathematics education specifically, dialogue-based ITS have shown significant promise. Xie et al. [17] developed a mathematics ITS for fractions that identifies student errors in real-time and provides diagnostic teaching, resulting in significant learning improvements compared to

control groups. Liu et al. [13] demonstrated that Chinese dialogue-based ITS for mathematics not only improved learning outcomes but also increased student motivation. These systems, however, typically rely on hand-crafted dialogue strategies and domain-specific engineering, limiting their scalability and generalizability.

5.2 Large Language Models in Education

The emergence of large language models has introduced new possibilities for educational applications [1, 4]. Recent models such as GPT-4 [1], Claude [2], and Gemini [7] demonstrate remarkable natural language understanding and generation capabilities, enabling more flexible and adaptive conversational interactions than traditional rule-based systems.

Several studies have explored LLMs' educational potential. Liu et al. [12] developed benchmarks for assessing LLMs' capabilities in evaluating multiple-choice questions in programming education, while Giannakos et al. [8] compared ChatGPT and human tutors in enhancing critical thinking skills through Socratic dialogue. Their findings suggest that while LLMs can engage in Socratic questioning, their effectiveness varies significantly depending on task complexity and pedagogical approach.

Recent work by Muñoz-Merino et al. [14] demonstrated that Socratic chatbots can enhance critical thinking in education, showing that students who interacted with Socratic tutors exhibited better reflection and analytical skills compared to those using standard chatbots. However, these studies typically focus on specific implementations rather than providing generalizable benchmarks for systematic evaluation.

5.3 Evaluation of Conversational AI Systems

Evaluating conversational AI systems, particularly in educational contexts, presents unique challenges [5]. Traditional metrics such as accuracy or BLEU scores fail to capture pedagogical quality. Recent approaches have explored LLM-as-a-judge methodologies [18], where advanced LLMs evaluate the outputs of other models based on specific criteria. This approach has shown promise in assessing dialogue quality, coherence, and helpfulness.

However, existing evaluation frameworks rarely address pedagogical considerations such as adherence to Socratic principles, scaffolding quality, or avoidance of direct answers. Our work builds upon these foundations by introducing domain-specific evaluation criteria tailored to Socratic teaching in mathematics education, combining automated student simulation with multi-dimensional assessment of pedagogical quality.

5.4 Socratic Method in AI-Assisted Education

The Socratic method has seen renewed interest in AI-assisted education [8, 14]. Recent implementations have demonstrated that AI systems can be designed to employ Socratic questioning strategies, guiding students toward discovery rather than providing direct answers. However, systematic evaluation of how well different AI models adhere to Socratic principles remains limited.

Our work distinguishes itself by providing a comprehensive, reproducible benchmark specifically designed to evaluate Socratic teaching capabilities across multiple dimensions, difficulty levels, and mathematical topics. This enables systematic comparison of different LLMs and identification of specific areas for improvement in AI-assisted Socratic pedagogy.

6. CONCLUSIONS

large language models' capabilities in Socratic teaching for mathematics education. Through systematic evaluation of three prominent LLMs across 1,000 carefully curated questions, we demonstrated that modern LLMs possess strong Socratic teaching capabilities, achieving overall scores exceeding 9.7/10 across all evaluation dimensions. Our multi-dimensional assessment revealed nuanced differences in pedagogical approaches, with Gemini 2.5 Flash leading in both overall performance (9.84) and teaching quality (9.61), while all models achieved near-perfect correctness (9.96).

Our framework provides researchers and practitioners with replicable infrastructure for assessing conversational AI tutoring systems, comprising a curated question dataset with rich annotations, an automated student simulator for scalable evaluation, and a multidimensional scoring system capturing pedagogical quality. The benchmark enables systematic comparison of different models and identification of specific areas for improvement.

Key findings include: (1) modern LLMs demonstrate strong baseline Socratic teaching capabilities with minimal task-specific training, (2) pedagogical effectiveness varies across topics and difficulty levels, suggesting opportunities for context-adaptive approaches, and (3) perfect mathematical correctness scores across all models indicate that concerns about AI tutors propagating errors may be less severe than anticipated for well-established mathematical domains.

Future work should expand the benchmark to even larger question sets (5000+ questions), incorporate diverse mathematical topics beyond those currently covered, validate automated evaluations against human expert judgments from experienced mathematics educators, enhance student simulator sophistication to capture authentic learning behaviors and emotional states, and conduct longitudinal field studies in authentic educational settings to measure real-world learning outcomes. As LLMs continue to advance, systematic benchmarks like ours will be crucial for ensuring these powerful technologies are deployed effectively and responsibly in educational contexts, ultimately enhancing learning outcomes for students worldwide.

The source code, dataset, and complete evaluation results will be made publicly available upon publication to enable reproducibility and encourage community engagement in advancing AI-assisted Socratic pedagogy.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023).

[2] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Anthropic Technical Report (2024).

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. 2021. On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv:2108.07258 (2021).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, Vol. 33. 1877 − 1901.

[5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology 15, 3 (2024), 1 − 45. doi:10.1145/3641289

[6] Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. 2001. Learning from human tutoring. Cognitive Science 25, 4 (2001), 471 − 533. doi:10.1207/s15516709cog2504_1

[7] Gemini Team, Google. 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2023).

[8] Michail Giannakos, Ioanna Voulgari, Sofia Papavlasopoulou, and Zacharoula Papamitsiou. 2024. Socratic wisdom in the age of AI: a comparative study of ChatGPT and human tutors in enhancing critical thinking skills. Frontiers in Education 10 (2024). doi:10.3389/feduc.2025.1528603

[9] Arthur C. Graesser, Sidney K. D'Mello, Xiangen Hu, Zhiqiang Cai, Andrew Olney, and Brent Morgan. 2011. AutoTutor. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 25. 1 − 7.

[10] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300 (2021).

[11] Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2015. The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. Cognitive Science 39, 4 (2015), 757 − 798. doi:10.1111/cogs.12184

[12] Ruihan Liu, Shuai Shi, Yuxuan Qi, Zhihao Zeng, Yichi Zhang, and Ge Li. 2024. A Benchmark for Testing the Capabilities of LLMs in Assessing the Quality of Multiple-choice Questions in Introductory Programming Education. In Proceedings of the 2024 ACM Virtual Global Computing Education Conference. 109 − 115. doi:10.1145/3649165.3690123

[13] Yi-Shan Liu, Sunny S. J. Lin, Chiung-Hui Chiu, and Tzu-Ying Wang. 2021. An Application of Chinese Dialogue-Based Intelligent Tutoring System in Remedial Instruction for Mathematics Learning. Educational Psychology 41, 2 (2021), 137 − 152. doi:10.1080/01443410.2020.1731427

[14] Pedro J. Muñoz-Merino, Raquel Gómez-Sánchez, et al. 2024. Enhancing Critical Thinking in Education by means of a Socratic Chatbot. arXiv:2409.05511 [cs.CY]

[15] Richard Paul. 1990. Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World. Center for Critical Thinking and Moral Critique (1990).

[16] Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. Educational Psychologist 46, 4 (2011), 197 − 221. doi:10.1080/00461520.2011.611369

[17] Jingqiao Xie, Rongqing Zhou, Junjie Luo, and Xiaoming Hu. 2023. Mathematics intelligent tutoring system for learning multiplication and division of fractions based on diagnostic teaching. Education and Information Technologies 28 (2023), 4883 − 4914. doi:10.1007/s10639-022-11553-z

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685 (2023).