Beyond Final Answers: Evaluating Large Language Models for Math Tutoring

Adit Gupta[1](✉), Jennifer Reddig[2], Tommaso Calò[3], Daniel Weitekamp[2], and Christopher J. MacLellan[2]

[1] Drexel University, 3230 Market Street, Philadelphia, PA, USA
adit.gupta@drexel.edu
[2] Georgia Institute of Technology, North Avenue, Atlanta, GA, USA
{jreddig,dweitekamp,cmaclellan}@gatech.edu
[3] Politecnico di Torino, Torino, Italy
tommaso.calo@polito.it

Abstract. Researchers have made notable progress in applying large language models (LLMs) to solve math problems, as demonstrated through efforts like GSM8k, ProofNet, AlphaGeometry, and Math-Odyssey. This progress has sparked interest in their potential use for tutoring students in mathematics. However, the reliability of LLMs in tutoring contexts—where correctness and instructional quality are crucial—remains underexplored. Moreover, LLM problem-solving capabilities may not necessarily translate into effective tutoring support for students. In this work, we present two novel approaches to evaluate the correctness and quality of LLMs in math tutoring contexts. The first approach uses an intelligent tutoring system for college algebra as a testbed to assess LLM problem-solving capabilities. We generate benchmark problems using the tutor, prompt multiple LLMs to solve them, and compare the solutions to those generated by the tutor. The second approach evaluates LLM as tutors rather than problem solvers. We employ human evaluators, who act as students seeking tutoring support from each LLM. We then assess the quality and correctness of the support provided by the LLMs via a qualitative coding process. We applied these methods to evaluate several ChatGPT models, including 3.5 Turbo, 4, 4o, o1-mini, and o1-preview. Our findings show that when used as problem solvers, LLMs generate correct final answers for 85.5% of the college algebra problems tested. When employed interactively as tutors, 90% of LLM dialogues show high-quality instructional support; however, many contain errors—only 56.6% are entirely correct. We conclude that, despite their potential, LLMs are not yet suitable as intelligent tutors for math without human oversight or additional mechanisms to ensure correctness and quality.

Keywords: Intelligent Tutors · Large Language Models · Generative AI · Math Education

A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15877, pp. 323–337, 2025.
https://doi.org/10.1007/978-3-031-98414-3_23

1 Introduction

Large language models (LLMs) have started to exhibit moderate proficiency at mathematical problem solving. For example, GPT-4 correctly solves over 90% of the problems in the GSM8K benchmark [5] and approximately 80% of the problems in the MATH benchmark [7] using advanced prompting techniques [22]. Although these results indicate progress, there are still many limitations. Findings from the GSM-Symbolic benchmark [20] suggest that LLMs struggle with perturbed or novel problem formulations that are easily solved by humans, indicating that their relatively high performance on standard benchmarks is partially due to memorization. Furthermore, LLM performance remains inconsistent across different problem classes, in contrast to traditional intelligent tutors, which provide 100% accurate support. These inconsistencies warrant a deeper investigation into the capabilities, limitations, and implications of LLMs for education.

Companies such as Duolingo and Khan Academy have started to leverage LLMs to offer personalized learning experiences, facilitate interactive problem-solving, and provide real-time feedback to learners. However, significant challenges remain to ensure the accuracy, reliability, and adaptability of LLMs in tutoring settings. Despite their remarkable capabilities, studies have shown that LLMs frequently produce plausible yet incorrect solutions to complex mathematical problems, especially in areas that require precise calculations and multi-step reasoning [11,20]. In mathematics, not only is the correctness of the final answer crucial, but also the quality of stepwise guidance that fosters effective learning. One recent classroom study comparing LLM-tutoring to traditional classroom instruction showed positive results [13], but another showed a negative result [4]. Considering that LLMs likely produce errors in around 10% of responses—using the best GSM8k performance as an optimistic measure—there is a possibility that they may do more harm than good. The manner in which LLMs confidently "hallucinate" incorrect yet seemly plausible information is a recipe for several possible negative effects [12]. In the best case, students' trust in LLM tutors may be eroded upon recognizing mistakes. At worst, LLM hallucinations may lead students to form misconceptions that compromise future learning.

LLM tutors mark an unusual inflection in the history of intelligent tutoring systems. It has been known for decades that automated computer-delivered tutors produce learning gains comparable to or greater than those of human tutors [17,26], who famously provide learning gains up to two standard deviations higher than those from traditional classroom instruction [2]. The original artificial intelligence (AI) tutors—hard-coded intelligent tutoring systems—found success through a cognitivist approach to tutoring: tracking and quantifying student knowledge by comparison to an expert model [6], and adapting instruction accordingly [21,25]. As compelling as LLMs' generative capabilities are, when used as standalone tutors, they arguably mark a regression in actual AI tutoring capabilities compared to traditional ITSs since they are consistently inaccurate and lack the cognition-oriented adaptivity of prior approaches.

This study evaluates the potential of LLMs in educational contexts by systematically assessing their performance on structured algebra tasks. We selected algebra for this study, given its long-standing use in previous research on ITSs [1,9,16]. This study aims to investigate the following research questions:

- **RQ1**: How accurately can LLMs generate solutions to the kinds of algebra problems currently supported by intelligent tutoring systems?
- **RQ2**: What is the accuracy and quality of the tutoring support provided by LLMs (e.g., scaffolding, hints, and feedback) on these algebra problems?

We employ two techniques to explore these questions: (1) an automated approach that uses an existing algebra tutor as a testbed for evaluating LLM problem solving and (2) a qualitative approach to assess the quality and correctness of LLM dialogues generated by having evaluators interactively prompt an LLM for tutor support. For the second method, we also conducted a thematic analysis [3] to identify and categorize observations about LLM tutoring behaviors.

The findings of this study contribute to research on ITS by providing empirical evidence on the strengths and limitations of LLMs in math tutoring contexts, thus enriching the ongoing discourse on the role of AI in supporting learning. Specifically, our study makes the following contributions:

- We introduce a novel method that uses intelligent tutors as testbeds for evaluating LLM problem solving.
- We introduce a second method for interactively evaluating LLM tutoring correctness and quality.
- We show that while LLMs largely generate responses aligned with pedagogical best practices, they frequently contain mistakes and inaccuracies, suggesting they are not yet ready for direct in-class deployment.
- We offer actionable guidelines for developers, emphasizing how LLMs can support aspects of tutoring, such as question generation and hint production, rather than serving as comprehensive, standalone tutoring solutions.

2 Related Works

Researchers have begun exploring the use of LLMs in educational applications. The existing literature shows that LLMs can generate worked examples and guide structured problem solving. For example, WorkedGen [28] uses prompt chaining and one-shot learning to produce interactive programming examples. Although user studies indicate that 77% of students found WorkedGen helpful, such self-reported feedback does not necessarily confirm improved learning outcomes. Similarly, Jamplate [15] harnesses AI-powered templates for idea generation, providing reflection-based scaffolding, but noting a tendency toward reduced critical thinking among students.

Although these studies highlight the potential of LLMs to create structured examples and facilitate reflective engagement, researchers must develop a consistent, stepwise evaluation framework for algebraic or multi-step reasoning tasks.

Existing benchmarks, such as GSM8K [5], evaluate only the final answer's accuracy, overlooking the intermediate steps and iterative feedback typical of most tutoring experiences [10]. As a result, there is still a need for a more systematic methodology that tests how effectively LLMs handle multi-step problems and adapt to the pedagogical requirements of a tutoring environment.

Another expanding line of work examines how LLMs function as tutors, focusing on the trade-off between personalized support and instructional accuracy for non-fluent English speakers. For example, a comparative study of models such as GPT-4, Llama-2-ko-DPO-13B, and eT5-chat reveals trade-offs between individualization and correctness [23]. Smaller models provided more personalized interactions, while GPT-4 exhibited greater correctness but less personalized assistance. Tutoring is an immensely personal activity, and both correctness and individualization are needed. These studies demonstrate the need for more investigation into LLMs' shortcomings in stepwise instruction and how to better integrate them into existing intelligent tutoring platforms. Current studies often prioritize correctness of the final answer, overlooking the quality of intermediate steps that are crucial for meaningful learning [27]. For example, in mathematics education, breaking problems down into their steps ensures that students grasp foundational concepts rather than simply arriving at the correct solution.

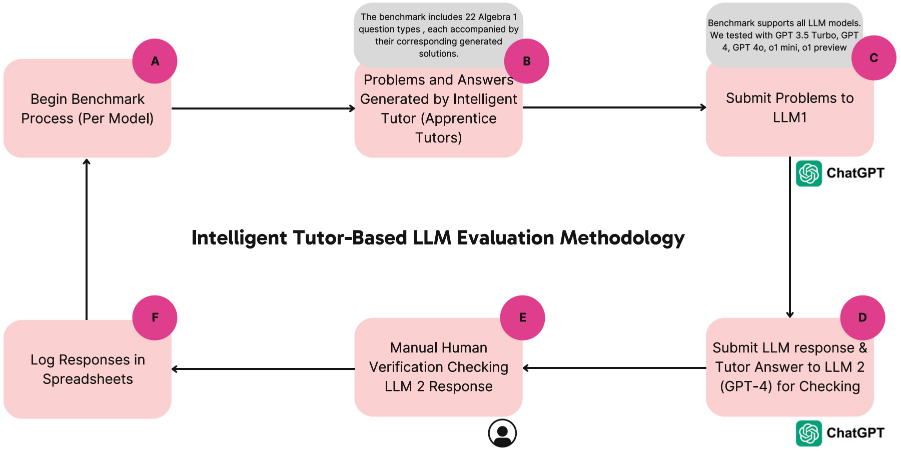

The benchmark includes 22 Algebra 1 question types , each accompanied by their corresponding generated solutions.
Benchmark supports all LLM models. We tested with GPT 3.5 Turbo, GPT 4, GPT 4o, o1 mini, o1 preview
A
Begin Benchmark Process (Per Model)
B
Problems and Answers Generated by Intelligent Tutor (Apprentice Tutors)
C
Submit Problems to LLM1
ChatGPT
Intelligent Tutor-Based LLM Evaluation Methodology
F
Log Responses in Spreadsheets
E
Manual Human Verification Checking LLM 2 Response
D
Submit LLM response & Tutor Answer to LLM 2 (GPT-4) for Checking
ChatGPT

Fig. 1. Our proposed intelligent tutor-based LLM evaluation process. The process begins with the testbed setup - hosted on Google Colab (**A**), followed by generating problems and solutions using the intelligent tutor (Apprentice Tutors, in our case) (**B**). For our evaluation, 22 types of problems are then submitted to LLM models such as GPT-3.5 Turbo and GPT-4 (**C**). Responses from each LLM are checked by submitting them along with the tutor answers to a second LLM (**D**). We then performed manual human verification to validate the accuracy of the second model's responses. Finally, all results are logged into a performance tracker spreadsheet (**F**).

The work in this paper aims to fill this gap by introducing a novel method that evaluates LLM performance on a wide range of math questions from college algebra, generated from the Apprentice Tutors platform [8]. This platform was designed as a web-based intelligent tutoring platform to support personalized learning in mathematics. The platform supports more than ten tutors covering topics like radicals, factoring polynomials, and solving logarithmic equations.

3 Methodology

We employ two complementary approaches to evaluate LLMs. First, we developed an automated approach that uses an existing intelligent tutoring system to assess LLM problem-solving accuracy. We generate problems from the tutor and submit them to multiple LLMs. We then use the tutor expert model to generate correct answers, which were then compared to the LLM answers using a second LLM (to account for minor variations in the math formatting). Second, to evaluate LLMs as tutors, we had evaluators interactively engage with the LLMs to request tutoring guidance as if they were students. We then qualitatively evaluated the tutor support generated by the LLMs.

We collected and analyzed data from both approaches to analyze the strengths and limitations of LLMs in structured problem-solving tasks. Figures 1 and 2 illustrate the workflows for these methodologies, which we describe below.

3.1 Evaluating LLM Using Intelligent Tutors as Testbeds

We developed our evaluation system in Python to automate tutor problem generation, LLM interaction, and response evaluation. This allowed us to systematically test multiple LLMs on a variety of educational tasks. For this study, we evaluated GPT-3.5 Turbo, GPT-4, GPT-4o, o1 Mini, and o1 Preview[1]. Although these models were the focus of this analysis, the benchmarking tool is designed to be extensible to other tutor content and can be easily adapted to test other models, such as Google's Gemini, Anthropic's Claude series, or Deepseek's open models.

The workflow for the tutor-based evaluation process is outlined in Fig. 1. The process begins with the testbed setup (**A**), where we define parameters for the evaluation, including the number and type of algebra problems to be tested. For this study, we identified 22 problem types from the Apprentice Tutors platform and generated five problems of each type. The problems and their corresponding step-by-step solutions were generated directly from the Apprentice Tutors software (**B**). The generated problems were then submitted to each LLM (**C**), which was prompted to produce a solution. The exact prompt provided to each LLM was:

[1] Model snapshots evaluated: `gpt-3.5-turbo-0125`, `gpt-4-0125-preview`, `gpt-4o-2024-05-13`, `o1-mini-2024-09-12`, and `o1-preview-2024-09-12`.

Math Problem Solving Prompt

You will be solving the math questions that are provided as strings. Your task is to parse each question, solve it step-by-step, and provide the final answer in LaTeX format.
Here are the math questions and their answers for verification: <question_answer_text>
Now, here are some new math questions that need answers: <next_question_text>
For each question, think through the <problem_type> problem step-by-step in ⟨scratchpad⟩ tags. Break down the problem into smaller sub-problems if necessary, and solve each one in a logical order. Show your work and reasoning at each step.
After you have thought through each problem and arrived at a final answer, confirm that it matches the provided answer in LaTeX format inside the corresponding ⟨answer⟩ tag.

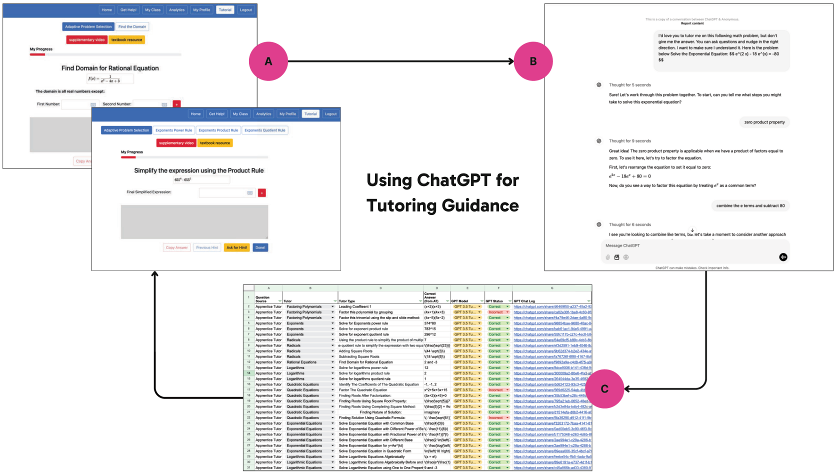
The benchmarking system processed the responses from the initial LLM prompt and extracted the outputs, generating a structured list of questions paired with their corresponding answers produced by the LLM. The LLM responses were then evaluated by submitting the correct solution (from the tutor) and the generated LLM solution to a second LLM (**D**) to verify accuracy and logical consistency. We used GPT-4 as the evaluation model in all tests. Each LLM was evaluated sequentially and the results were recorded and analyzed before proceeding to the next model. The exact prompt used by the second LLM to evaluate each answer was:

LLM Evaluation Prompt

Just say True or False (nothing else): does <LLM_generated_response> equal the same as <ground_truth_response_from_tutor>?

To further validate the quality of the LLM evaluation, we performed manual human verification (**E**). During this process, reviewers compared the ground-truth responses generated by the Apprentice Tutors platform with the responses produced by the LLM and the correctness assessments provided by the second LLM. In certain cases, discrepancies arose due to differences in interpretation, such as when the second LLM marked an expression like $\sqrt{4}$ as incorrect because it expected the simplified answer of 2. These instances were noted and the human reviewer marked the answer as correct if it was mathematically accurate in its final form. However, stepwise solutions were also considered, ensuring that intermediate simplifications (e.g., distinguishing between $\sqrt{4}$ and 2 when necessary) aligned with expected problem-solving conventions.

Finally, all data collected, including LLM responses, human evaluations, and any discrepancies identified during the validation process, were systematically recorded in a performance tracker spreadsheet (**F**). This structured logging approach facilitated detailed analysis and allowed for a robust comparison between different LLM models and problem types. The goal was to gain insights into their performance and limitations.



Find Domain for Rational Equation

$f(x) = \frac{1}{x^2 - 4x + 3}$

The domain is all real numbers except:

First Number: Second Number:

Simplify the expression using the Product Rule

$655^8 \cdot 655^5$

Final Simplified Expression:

This is a copy of a conversation between ChatGPT & Anonymous.
Report content

I'd love you to tutor me on this following math problem, but don't give me the answer. You can ask questions and nudge in the right direction. I want to make sure I understand it. Here is the problem below Solve the Exponential Equation: $$ e^{2 x} - 18 e^{x} = -80 $$

Thought for 5 seconds

Sure! Let's work through this problem together. To start, can you tell me what steps you might take to solve this exponential equation?

zero product property

Thought for 9 seconds

Great idea! The zero product property is applicable when we have a product of factors equal to zero. To use it here, let's try to factor the equation.

First, let's rearrange the equation to set it equal to zero:

$e^{2x} - 18e^x + 80 = 0$

Now, do you see a way to factor this equation by treating e^x as a common term?

combine the e terms and subtract 80

Thought for 6 seconds

I see you're looking to combine like terms, but let's take a moment to consider another approach

ChatGPT can make mistakes. Check important info.

	A	B	C	D	E	F	G
1	Question Source	Tutor	Tutor Type	Correct Answer (from AT)	GPT Model	GPT Status	GPT Chat Log
2	Apprentice Tutor	Factoring Polynomials	Leading Coeffieent 1	(x+2)(x+3)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/d6469f55-a237-45a2-92
3	Apprentice Tutor	Factoring Polynomials	Factor this polynomial by grouping	(4x+1)(4x+3)	GPT 3.5 Tu...	Incorrect	https://chatgpt.com/share/ca02e30f-1be8-4c63-9f
4	Apprentice Tutor	Factoring Polynomials	Factor this trinomial using the slip and slide method:	(4x−5)(5x−2)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/f4a79e46-2dae-4a80-9c
5	Apprentice Tutor	Exponents	Solve for Exponents power rule	374^80	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/98854bae-9680-40ac-8
6	Apprentice Tutor	Exponents	Solve for exponent product rule	783^15	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/bab61ac1-94e5-4991-a
7	Apprentice Tutor	Exponents	Solve for exponent quotient rule	296^12	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/55fc117b-c27c-4ec0-b8
8	Apprentice Tutor	Radicals	Using the product rule to simplify the product of multip	7	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/64e69cf5-b89c-4cb3-8b
9	Apprentice Tutor	Radicals	e quotient rule to simplify the expression with two squa	\(\frac{\sqrt{23}}{	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/ef3d2591-1eb9-4046-8c
10	Apprentice Tutor	Radicals	Adding Square Roots	\(44 \sqrt{3}\)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/9b62d374-b2e2-434e-a
11	Apprentice Tutor	Radicals	Subtracting Square Roots	\(18 \sqrt{5}\)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/fa76726f-6f85-4167-8b
12	Apprentice Tutor	Rational Equations	Find Domain for Rational Equation	2 and -3	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/f9892a9a-c4d8-4f75-a9
13	Apprentice Tutor	Logarithms	Solve for logarithms power rule	12	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/8dce9006-b141-438d-9
14	Apprentice Tutor	Logarithms	Solve for logarithms product rule	2	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/300009a2-80e6-4fa3-a
15	Apprentice Tutor	Logarithms	Solve for logarithms quotient rule	1	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/264044da-3e35-4682
16	Apprentice Tutor	Quadratic Equations	Identify The Coefficients of The Quadratic Equation	-1, -1, 2	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/dd624123-93c3-425
17	Apprentice Tutor	Quadratic Equations	Factor The Quadratic Equation	x^2+5x+3x+15	GPT 3.5 Tu...	Incorrect	https://chatgpt.com/share/569d6225-54eb-4fd
18	Apprentice Tutor	Quadratic Equations	Finding Roots After Factorization:	(5x+2)(x+5)=0	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/35b53bef-c28c-446
19	Apprentice Tutor	Quadratic Equations	Finding Roots Using Square Root Property:	\(\frac{\sqrt{5}}{7	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/795a21eb-0652-48ed
20	Apprentice Tutor	Quadratic Equations	Finding Roots Using Completing Square Method:	\(\frac{5}{2} + \fra	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/b2d3e84a-bdb4-482c-a
21	Apprentice Tutor	Quadratic Equations	Finding Nature of Solution:	imaginary	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/d1514efa-d8b2-4416-a
22	Apprentice Tutor	Quadratic Equations	Finding Solution Using Quadratic Formula:	\(- \frac{\sqrt{61}	GPT 3.5 Tu...	Incorrect	https://chatgpt.com/share/5fa36290-d912-41f1-9b
23	Apprentice Tutor	Exponential Equations	Solve Exponential Equation with Common Base	\(\frac{4}{3}\)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/f3203172-7baa-4141-8
24	Apprentice Tutor	Exponential Equations	Solve Exponential Equation with Different Power of Ba	\(- \frac{11}{8}\)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/0ad00eb5-3c90-46f3-9
25	Apprentice Tutor	Exponential Equations	Solve Exponential Equation with Fractional Power of E	\(- \frac{41}{7}\)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/b1175348-e263-4d6b-8
26	Apprentice Tutor	Exponential Equations	Solve Exponential Equation with Different Base	\(\frac{2 \ln{\left(	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/2ae994e1-c29a-4286-b
27	Apprentice Tutor	Exponential Equations	Solve Exponential Equation for y=Ae^(kt)	\(- \frac{\log{\left(	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/2ae994e1-c29a-4286-b
28	Apprentice Tutor	Exponential Equations	Solve Exponential Equation in Quadratic Form	\ln{\left(10 \right)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/89eaa006-35cf-4bcf-a7
29	Apprentice Tutor	Logarithmic Equations	Solve Logarithmic Equations Algebraically	\(x = e\)	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/feebe54c-ffb5-4ada-9a
30	Apprentice Tutor	Logarithmic Equations	Solve Logarithmic Equations Algebraically Before and	\(\frac{e^{\frac{1}	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/89e6191a-e737-4d1f-b
31	Apprentice Tutor	Logarithmic Equations	Solve Logarithmic Equation using One to One Propert	9 and -3	GPT 3.5 Tu...	Correct	https://chatgpt.com/share/c45a866b-ad33-4380-9

Fig. 2. Our process for evaluating an LLM via interactive prompting. This diagram illustrates the process using ChatGPT's chat interface. The workflow begins by having evaluators prompt ChatGPT to provide tutor guidance on tutor problems as if they were students (**A**). Evaluators interactively submit queries and receive step-by-step guidance from ChatGPT (**B**). The evaluators systematically log each chat dialogue in a spreadsheet tracker for analysis (**C**).

3.2 Evaluating LLMs via Interactive Prompting

We conducted a second study to assess how the LLMs perform when interacting with learners as tutors (see Fig. 2). This study was designed to provide a qualitative perspective on the educational capabilities of the models, in contrast to the previous automated evaluation of their problem-solving capabilities. We performed manual evaluation of the LLMs using a standardized variation of the prompt from Salman Khan's widely cited ChatGPT interview [14] to ensure consistency between sessions. By comparing these human-guided interactions with the outputs of the intelligent tutor, we investigated how well the step-by-step guidance of LLMs align with real user queries and misconceptions in a math tutoring context. Here is the prompt that was used:

Interactive Tutoring Prompt

I'd love you to tutor me on this following math problem, but don't give me the answer. You can ask questions and nudge me in the right direction. I want to make sure I understand it. Here is the problem below. <Problem>

The problems were taken directly from the Apprentice Tutors and entered into ChatGPT. Three evaluators interacted with the model as it guided them through problem-solving, responding to ChatGPT's hints and prompts as they progressed. After each session, they logged a link to the chat history, the final answers provided by ChatGPT, and whether the responses matched the correct answers generated by the Apprentice Tutors. An example of this recording is shown in part B of Fig. 2.

After collecting all responses, two independent reviewers assessed the interactions to answer the following questions about each tutoring dialogue:

- **Quality:** Do the steps represent a high-quality tutoring interaction?
- **Correctness:** Were all the LLM responses in the dialogue correct?

To answer the first question and classify response **quality**, each dialogue was evaluated with respect to a structured rubric that scored clarity of explanation, feedback, scaffold support, problem-solving strategy, and encouragement and reinforcement.[2] Each criterion was scored on a scale from 1 to 4. The rubric was designed to measure several key aspects of tutoring quality. This structured approach aimed to reflect established learning science principles. We summed the scores across all criteria. If the total score was 10 or below, the response was categorized as "No" (not good quality); if the score was above 10, it was categorized as "Yes" (good quality). Once the scores were converted into Yes/No labels, we measured inter-rater reliability using Cohen's Kappa [19] to assess agreement between reviewers and confirm the robustness of our classifications.

To answer the second question and assess response **correctness**, each reviewer also independently assessed whether all the LLM-generated content was correct. These evaluations consisted of considering each LLM response from the dialogue, and noting any mistakes or errors. If there were any errors, then the dialogue was coded as incorrect, otherwise it was recorded as correct. Similar to question 1, we measured inter-rater reliability of the two evaluations using Cohen's Kappa. We also conducted a thematic analysis [3] of the LLM responses to identify recurring patterns in tutoring interactions. The goal was to identify and categorize common patterns, counting their frequency and noting whether they corresponded to positive or negative tutoring behaviors.

4 Results and Analysis

We present the results of the two evaluation methods used to assess the performance of LLMs in math tutoring contexts. We first present the results from our tutor-based evaluation method and then the results from evaluators interactively prompting the LLMs as if they were students.

[2] The rubric is available here: https://osf.io/k4wqc.

4.1 Tutor-Based LLM Evaluation Results

Table 1 summarizes the results from the tutor-based evaluation. The apprentice tutors had 22 problem types and we sampled 5 problems of each type to produce a test set that contained 110 problem and answer pairs.

Table 1. The tutor-based LLM evaluation results, which only assessed the final answer.

Model	# Problem Types	# Problems	# Correct	Accuracy
GPT-3.5 Turbo	22	110	85	77.3%
GPT-4	22	110	83	74.5%
GPT-4o	22	110	107	97.3%
o1-mini	22	110	101	91.8%
o1-preview	22	110	94	85.5%
Overall Avg.	22	110	94	85.5%

Table 2. Assessments of LLM tutoring interaction quality and accuracy across 30 problems. The columns show final answer accuracy as well as the percentage of LLM dialogues that were classified as high quality and fully correct, as indicated by reviewers R1 and R2. The number of problems in each case is noted in parentheses.

Model	Final Accuracy	% High Quality		% Fully Correct	
		R1	R2	R1	R2
GPT 3.5 Turbo	90.0% (27)	90.0% (27)	90.0% (27)	46.7% (14)	53.3% (16)
GPT 4	83.3% (25)	93.3% (28)	93.3% (28)	43.3% (13)	50.0% (15)
GPT 4o	93.3% (28)	90.0% (27)	90.0% (27)	70.0% (21)	80.0% (24)
o1 mini	86.7% (26)	86.7% (26)	80.0% (24)	56.7% (17)	43.3% (13)
o1 preview	90.0% (27)	90.0% (27)	96.7% (29)	73.3% (22)	50.0% (15)
Overall Avg.	88.6%	90.0%		56.6%	

We identified twenty-five instances (6.3% of total responses) where the second LLM marked answers incorrectly. From our observations, the second LLM would incorrectly mark the answer when comparing the tutor-generated response to the LLM-generated response for the following reasons: **(1)** a mismatch in the operational order (e.g., $(3+6)\times 2$ vs. $3+(6\times 2)$), **(2)** differences in simplification (e.g., $\frac{2}{4}$ vs. 0.5), and **(3)** differences in operator context (e.g., multiplication represented by "x" vs. "*").

4.2 Interactive Prompting-Based LLM Evaluation

The evaluators prompted each of the five models to provide tutoring support on the same set of 30 problems, resulting in a total of 150 LLM dialogues. The two reviewers independently analyzed each dialogue to classify whether it was high quality (according to the rubric) and fully correct. We also evaluated the final answer accuracy from each dialogue by comparing it to the tutor solution. Table 2 shows the results of these assessments, breaking out the accuracy of final LLM answers alongside reviewer assessments of the quality and correctness of the LLM dialogues. We calculated Cohen's Kappa (κ) to evaluate the reviewer agreement for both the quality and correctness ratings. This score, which ranges from 0 to 1, represents agreement after correcting for chance. A score greater than 0.7 is typically viewed as strong agreement. For the quality ratings, Cohen's $\kappa_{Quality} \approx 0.85$, and for the assessment of whether the LLM dialogues were entirely correct, Cohen's $\kappa_{Correctness} \approx 0.82$. These scores indicate very strong agreement between the independent reviewers.

Finally, the reviewers evaluated each model's performance, documenting key behavioral patterns and noting any common issues. Table 3 summarizes these findings, classifying observations as positive or negative based on their potential impact on learners. This analysis highlights the strengths and weaknesses of each model's tutoring approach, offering insight into their effectiveness in guiding students through problem-solving.

5 Discussion

Both of our evaluation methods suggest that the LLMs show reasonable final answer accuracy. Our tutor-based evaluation showed that GPT-4o had the highest final-answer accuracy at 97.3%. In the interactive prompting-based evaluation, we found that GPT-4o also got the highest accuracy at 93.3%. Although these accuracies seem reasonable, it still means that these models will generate incorrect final answers for about 1 in 18 problems. Surprisingly, GPT-4 ranked last, and the model ranking changed based on the evaluation.

Newer models (o1-mini) often performed worse than older ones (GPT-4o), despite expected improvements. This suggests that LLM performance can no longer be expected to improve with each new model release. Also, there are still gaps in how LLMs process multi-step math questions, even with chain-of-though models. In terms of final answer accuracy, we observed that the interactive prompt-based evaluation results were higher than those from the tutor-based evaluation, probably because human testers engaged in multi-turn interactions, allowing LLMs to refine responses. In contrast, the tutor-based evaluation provided only a single prompt, requiring the model to solve problems correctly in one step.

During our interactive prompting-based evaluation, we found that LLMs generate high-quality tutoring support most of the time. Although GPT-4 had the lowest final answer accuracy, it scored near the top in terms of quality, with 93.3% of its chat dialogues being classified as having high pedagogical quality.

Table 3. Summary of key observations from the interactive tutoring evaluation.

Observation	Occurrences	Sentiment
The final answer was correct, even though there were Inconsistencies in the sub-steps	6	Negative
Even though the prompt was not to share an answer, it was possible to obtain the answer by manipulating responses (via yes/no questions)	4	Negative
For topics like factoring, there was an overemphasis on teaching basics (e.g., multiplying) instead of demonstrating specific methods (e.g., "slip and slide")	4	Negative
LLM over-indexes on ensuring the final answer is correct rather than emphasizing the student's step-by-step skill acquisition	3	Negative
LLM occasionally produces an incorrect conclusion and refuses to accept a correct student answer	4	Negative
Sub-steps are sometimes flagged as incorrect even though they are actually correct	3	Negative
Difficult math notation (e.g., quadratic expressions) can be challenging to input from a standard keyboard	2	Negative
LLM is flexible about answer formats, accepting multiple notational styles	3	Positive
LLM excels at generating hints and extra worked examples to support instruction	2	Positive
LLM provides encouraging feedback and positive reinforcement, which could benefit student well-being	7	Positive
LLM nudges students to answer queries in sequence when they attempt to skip ahead	2	Positive

Although the LLMs achieved reasonable final answer accuracies, we found that their full tutor dialogues often contained mistakes. Along this dimension, GPT-4o achieved the best performance, with 75% of its dialogues being classified as fully correct (averaging across the two reviewers). Across all five of the LLMs, only 56.6% of the tutor dialogues were entirely correct. This suggests that in almost half of interactive LLM tutoring sessions, students will receive partially incorrect instruction. These results suggest that LLM error rates at tutoring tasks are likely much higher than final answer-based benchmarks suggest. This raises concerns about their use as standalone tutors, as less-than-perfect accuracy can harm learners. If one in two dialogues contains errors, students may lose trust in the tutor and, worse, develop misconceptions that hinder future learning. Our results also suggest that future evaluations must consider the correctness of the entire tutoring dialogue, not just the final answer accuracy.

Traditional intelligent tutoring systems employ explicit scaffolding, guiding students through sequenced substeps that stem from research-based design methods such as cognitive task analysis [18]. In contrast, large language models often provide a complete solution or a fixed step-by-step explanation, offering limited adaptive support for the intermediate reasoning of learners. For example, in factoring problems, LLMs often provide overly general guidance, e.g. generic multiplication rules, instead of the specific "slip and slide" strategy explicitly requested in the task. Although such responses may earn full credit in automated evaluations, omitting the specified method diminishes the tutoring quality.

Table 3 summarizes reviewer observations of tutoring interactions. Reviewers noted that the LLMs sometimes refused to accept correct answers, miscalculated sub-steps, or overemphasized fundamentals at the expense of specialized techniques. However, the LLMs also provided flexible response formats, detailed hints, and encouraging feedback. Manual review of chat logs revealed inconsistencies in how LLMs handled intermediate steps. While they often produced correct final answers, sub-steps were occasionally miscalculated or erroneously flagged as incorrect (8 instances, as shown in Table 3). These errors fell into three categories: (1) simplified vs. unsimplified answers (e.g., 2 vs. $\sqrt{4}$), (2) differences in term order (e.g., $\sqrt{3} + \sqrt{4}$ vs. $\sqrt{4} + \sqrt{3}$), and (3) formatting mismatches (e.g., missing required LaTeX tags). These issues highlight inconsistencies in how LLMs evaluate semantic equivalence.

Although LLMs may be incorrect and pedagogically misaligned compared to ITS when tutoring, recent work demonstrates how LLMs could support ITS in hint generation [24]. By leveraging the expert model of ITS, LLMs can generate correctness feedback personalized to student responses without needing the LLM to perform any mathematical calculations [24]. If integrated with other educational technologies, they could offer several potential benefits. Their ability to generate hints, provide alternative explanations, and accommodate various answer formats makes them flexible and adaptive. However, our finding that LLMs sometimes mark correct answers as incorrect—even when provided with the solutions—suggest that LLMs integrations will need to be carefully evaluated before deployment.

Additionally, their use of positive reinforcement—such as motivational nudges and encouraging feedback—could help foster an engaging learning environment, provided that motivational support is offered equitably to all learners. For example, several chat logs included statements like, "Way to go, you are close to the answer!" or "That's not right, but let's keep trying." These reinforcements might help motivate learners to persist and promote sustained engagement with educational tools. This approach aligns with many learners' needs for constructive feedback and encouragement [8]. Future research should systematically evaluate the motivational potential of LLMs' interactions and whether they translate into improved learning outcomes.

6 Limitations and Future Work

One limitation of our LLM evaluation methods is that they were conducted as an offline evaluation instead of with students. We chose our approach because we knew that LLMs have reliability issues and we did not want to cause harm to students by giving them incorrect tutoring guidance during our experiments. Although our approach provides a means of safe, controlled evaluation, it may not fully capture the unique ways in which real students would engage with LLM tutors. A future iteration of this work could involve deploying the system in real-world educational settings and analyzing chat logs generated from authentic student interactions to gain a more comprehensive understanding of LLM performance. However, since our work suggests that LLMs make mistakes in just over half of their student tutoring dialogues, future research should try to mitigate the risks that LLMs pose to students.

Furthermore, this study was conducted using an earlier version of the Apprentice Tutors platform, which focused exclusively on math-related questions. The Apprentice Tutors platform has since been expanded to include other types of questions, such as those related to nursing education. Future research could explore how LLMs perform in this and other domains, extending the scope of evaluation to understand their domain-specific adaptability and effectiveness.

Finally, another limitation is that the analysis presented was restricted to a set of LLMs within the ChatGPT family. With the rapid development of new LLMs, such as Google's Gemini, Anthropic's Claude and several open-source LLMs, there is an opportunity to expand this study to evaluate these emerging models too. Comparing performance across a broader range of LLMs would provide a more holistic view of their strengths and weaknesses in educational contexts. In addition, this work does not thoroughly address potential biases and limitations in LLM-generated tutoring content—a critical issue for ensuring safe and effective educational use. Lastly, this study used the publicly accessible versions of the ChatGPT models. In practice, commercial production environments often deploy models that are fine-tuned to specific domains or tasks. Evaluating a custom-tuned LLM tailored to specific educational needs could offer a more accurate view of how these tools would perform in real-world applications.

7 Conclusion

In this study, we evaluated the ability of various LLMs to solve college algebra problems and to interactively provide step-by-step tutor guidance. We evaluated multiple models, including GPT-3.5 Turbo, GPT-4, GPT-4o, o1 Mini and o1 Preview, identifying both their strengths and limitations. The results presented in this study, though commendable, are significantly lower than the 100% accuracy achieved by traditional intelligent tutors on the same set of problems. While we saw an overall final accuracy of 85.5% with the automated tutor-based evaluation and 88.6% with our interactive prompting-based evaluation, our analysis of the entire LLM tutoring dialogues showed that only 56.6% were entirely correct. This discrepancy suggests a core limitation of using LLMs as tutors: while they often generate correct final answers, ensuring the pedagogical soundness and accuracy of intermediate steps remains challenging.

Despite these limitations, LLMs exhibit several capabilities that have the potential to improve learning outcomes. Their flexibility in accepting diverse answer formats, the ability to generate hints and alternative problem explanations, and the use of positive reinforcement, such as motivational nudges, could help foster a more supportive and engaging learning environment. However, there are risks associated with the deployment of LLMs in educational settings. For example, biases within the models may lead to inflexibility in pedagogical approaches, such as internal biases that favor some methods of solving problems over others. Furthermore, inaccuracies in responses—with around one in two dialogues containing errors—can undermine the trust of students in the guidance of the tutor and reduce their confidence in the system. To address these challenges, future work might explore how to leverage their independent capabilities, such as problem generation, hint generation, and positive reinforcement. By balancing these strengths with strategies to manage and mitigate their limitations, LLMs could effectively supplement other educational technologies, such as intelligent tutoring systems.

Acknowledgments. Funded in part by NSF AI Institutes program (#2247790 and #2112532). Views are the author's and do not reflect those of the U.S. Govt. We also thanks Kaitlyn Crutcher and Momin Siddiqui for their feedback and suggestions.

References

1. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: The cognitive tutor authoring tools (CTAT): preliminary evaluation of efficiency gains. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 61–70. Springer, Heidelberg (2006). https://doi.org/10.1007/11774303_7
2. Bloom, B.S.: The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educ. Res. **13**(6), 4–16 (1984)
3. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)
4. Cicchetti, C.C.: AI in higher education: does not help, might hurt. J. Educ. Bus. **99**(7–8), 438–443 (2024)
5. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, J.H., Raffel, C., Schulman, J.: Training verifiers to solve math word problems. arXiv:2110.14168 (2021)
6. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adapted Interact. (1994)
7. Fang, M., Wan, X., Lu, F., Xing, F., Zou, K.: MathOdyssey: benchmarking mathematical problem-solving skills in large language models using odyssey math data. arXiv:2406.18321 (2024)
8. Gupta, A., Siddiqui, M., Smith, G., Reddig, J., MacLellan, C.: Intelligent tutors for adult learners: an analysis of needs and challenges. arXiv:2412.04477 (2025)
9. Heffernan, N., Heffernan, C.: The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Int. J. Artif. Intell. Educ. **24**, 470–497 (2014)
10. Heffernan, N.T., Koedinger, K.R., Razzaq, L.: Expanding the model-tracing architecture: a 3rd generation intelligent tutor for algebra symbolization. Int. J. Artif. Intell. Educ. **18**(2), 153–178 (2008)

11. Hendrycks, D., et al.: Measuring mathematical problem solving with the MATH dataset. arXiv:2103.03874 (2021)
12. Huang, L., et al.: A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv:2311.05232 (2023)
13. Kestin, G., Miller, K., Klales, A., et al.: AI tutoring outperforms active learning. Res. Square (2024). https://doi.org/10.21203/rs.3.rs-4243877/v1
14. Khan, S.: Sal Khan explores ChatGPT in education. YouTube video. Published on Khan Academy's Official Channel, March 2023
15. Kim, J.S.: Jamplate: exploring LLM-enhanced templates for idea reflection. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 1–14 (2023)
16. Koedinger, K.R., Anderson, J.R., Hadley, W.H., Mark, M.A.: Intelligent tutoring goes to school in the big city. Int. J. Artif. Intell. Educ. **8**, 30–43 (1997). ffhal-00197383
17. Kulik, J.A., Fletcher, J.: Effectiveness of intelligent tutoring systems: a meta-analytic review. Rev. Educ. Res. **86**(1), 42–78 (2016)
18. Lovett, M.C.: Cognitive task analysis in service of intelligent tutoring systems design: a case study in statistics. In: Goettl, B.P., Halff, H.M., Redfield, C.L., Shute, V.J. (eds.) Intelligent Tutoring Systems. LNCS, vol. 1452, pp. 234–243. Springer, New York (1998)
19. McHugh, M.L.: Interrater reliability: the Kappa statistic. Biochemia Medica **22**(3), 276–282 (2012)
20. Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., Farajtabar, M.: GSM-symbolic: understanding the limitations of mathematical reasoning in large language models. arXiv:2410.05229 (2024)
21. Nwana, H.S.: Intelligent tutoring systems: an overview. Artif. Intell. Rev. **4**(4), 251–277 (1990)
22. OpenAI. GPT-4 technical report. arXiv:2303.08774 (2023)
23. Park, J.W., Kim, M.J., Lee, S.W.: Developing conversational intelligent tutoring for speaking skills in foreign language education. In: Koch, F., Zhang, L., Chen, F., Dillenbourg, P. (eds.) Artificial Intelligence in Education, pp. 123–134. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-63028-6_11
24. Reddig, J.M., Arora, A., MacLellan, C.: Generating in-context, personalized feedback for intelligent tutors with large language models (2024)
25. VanLehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Educ., 227–265 (2006)
26. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. (2011)
27. Xia, S., et al.: Evaluating mathematical reasoning beyond accuracy. arXiv:2404.05692 (2024)
28. Zhang, Y., et al.: Workedgen: using LLMs to generate interactive worked programming examples. In: Proceedings of the 2023 Conference on Learning at Scale, pp. 123–134 (2023)