LLM Based Math Tutoring: Challenges and Dataset

Pepper Miller, Kristen Dicerbo
Khan Academy

Abstract

Large Language Models (LLMs) face documented challenges in solving mathematical problems. While substantial work has been done to quantify and improve LLMs' abilities to solve static math problems, evaluating their performance in real-time math tutoring scenarios presents distinct challenges that remain underexplored. This paper specifically addresses the accuracy of LLMs in performing math correctly while tutoring students. It highlights the unique difficulties of this context, classifies types of interactions students may have with an LLM, presents a dataset, Conversation-Based Math Tutoring Accuracy Dataset (CoMTA Dataset), for evaluating the mathematical accuracy in tutoring scenarios, and discusses techniques to address these issues. Additionally, it evaluates the mathematical accuracy of a range of models in LLM-based tutoring.

1 Introduction

LLMs are increasingly utilized in educational settings, particularly for tutoring [1]. However, the act of using an LLM as a tutor to help students with their math difficulties is a unique problem that has yet to be fully investigated. Unlike solving math problems directly, tutoring requires the LLM to track student progress through a problem or series of problems and guide the student's thought process without simply providing answers. This involves inferring the steps the student has performed, evaluating those steps against the correct methodology (or methodologies as there can be multiple solution paths), and identifying any mistakes. Furthermore, the tutor must discern which step the student is currently on and navigate various subtleties in the student's approach and understanding. These challenges make the task of using LLMs for math tutoring particularly complex. Existing benchmark datasets such as MATH and GSM8K have focused primarily on evaluating LLMs' ability to solve math problems accurately [2, 3]. While these datasets provide valuable insights, they do not address the complexities of using LLMs in a tutoring context. This paper addresses these challenges and proposes a benchmark dataset to evaluate LLM performance in tutoring scenarios.

2 Problem Definition

Tutoring in math differs significantly from solving math problems. The primary goal of a tutor is to help the student understand and solve problems independently, rather than just providing the correct answer. This distinction brings numerous challenges:

- **Guiding the learning process without giving away solutions:** Effective tutoring involves helping students find the answers themselves, which can be difficult for LLMs that are optimized to provide direct answers.
- **Understanding the student's problem-solving process is crucial:** The tutor needs to track the student's progress, identify misunderstandings, and provide targeted guidance, requiring the LLM to have a nuanced understanding of the student's approach and where they might be going wrong.
- **Managing multiple and shifting problems during a tutoring session:** Students might introduce new problems or switch between problems during a session. The LLM needs to adapt to these shifts and keep track of different problems simultaneously, as human tutors do.
- **Handling ambiguity in students' language and maintaining engagement and motivation:** Students often express themselves ambiguously or unclearly. The LLM must interpret these inputs correctly and maintain the student's engagement and motivation throughout the tutoring session.
- **Resisting students' attempts to extract answers from the LLM through various means:** In a tutoring context, students might try to get the answers from the LLM through direct requests, repeated questions, or by asking for step-by-step assistance without truly engaging with the material. The LLM must navigate these interactions to provide support without simply giving away the answers.

3 Challenges

3.1 Student Behavior

- **Asking for the Answer:** Students may directly request answers, once or repeatedly, which the tutor must avoid giving. This requires the LLM to redirect the student's focus towards problem-solving techniques rather than simply providing the answer.
- **Interpreting Incorrect Answers:** Students may repeatedly provide wrong answers hoping the tutor will eventually reveal the correct one.

However, these repeated wrong answers could also be a result of the student being lost, and the tutor needs to understand which scenario is occurring.

3.2 Tutor's Understanding

- **Student's Process:** The tutor must understand the student's approach and progress in solving the problem. This includes recognizing the student's current step in the problem-solving process and any potential errors they might have made.
- **Shifting Problems:** Recognizing when the focus shifts between different problems or sub-problems. The LLM must keep track of the original problem, any introduced sub-problems, and the context of the student's questions and answers.
- **Language Ambiguity:** Parsing unclear or ambiguous student inputs effectively. The LLM must be able to interpret vague or poorly phrased questions and guide the student towards clarity.

3.3 Tutor's Response

- **Providing Elaborated Feedback:** Providing an accurate indication of the correctness of the student's response. The LLM must provide the appropriate level of information about why the student's response is or is not correct [4].
- **Tutoring Moves:** Providing the right next response from a large range of possibilities, from summarizing to pushing for more explanation, that will promote learning [5].
- **Contextually Appropriate:** Using grade level appropriate concepts and not invoking more advanced material than necessary.
- **Motivation and Support:** Balancing engagement and the level of support to keep the student motivated. The LLM must provide enough support and hints to keep the student motivated without making the problem-solving process too easy or too difficult [6].

3.4 LLM-Specific Issues

- **Subtle Mistakes:** One of the challenges for LLMs in a tutoring context is identifying and correcting answers that are nearly correct. These subtle mistakes require the LLM to not only recognize the error but also understand the correct solution to guide the student appropriately.
- **Fixation on Errors:** Another issue is the fixation on errors, where the LLM might focus on a mistake made by the student, even after the student

has corrected it. The tutor must be able to recognize when a student has corrected a mistake and shift the focus to further learning rather than repeatedly addressing the past error.

- **Maintain the Problem:** When a problem is a near variant of common or more straightforward problems, the LLM can shift to answering and tutoring that more common problem. It can lose context of the variation, particularly in longer conversations.

- **Taking Student's Word:** LLMs must be capable of challenging incorrect assertions made by students. If a student incorrectly insists that their answer is correct, the LLM must be able to effectively question and guide the student towards the correct understanding, rather than simply accepting the incorrect information [7, 8].

4 Benchmark Dataset

The many challenges outlined above demonstrate the complexity of tutoring. In our piloting of Khanmigo we observed a fundamental problem; the model sometimes provides incorrect evaluation information to the student. That is, it indicates the student is correct when they are wrong or incorrect when they are right. This is not merely an artifact of the model incorrectly computing an answer, but is the result of many of the other factors described above. We propose a benchmark dataset designed to evaluate the accuracy of mathematical response evaluation of LLMs in tutoring scenarios. This dataset comprises 188 dialogues between an LLM tutor and a student. Each conversation is truncated at the point where the student makes a mathematical claim. Some of these claims are correct, while others are incorrect. This dataset can be used to assess an LLM's ability to evaluate math in the tutoring context by having the LLM generate a response to the student's claim and determining if it correctly identifies mistakes or accurately acknowledges correct statements.

The full CoMTA Dataset can be found at at https://github.com/Khan/tutoring-accuracy-dataset.

4.1 Dataset Composition

- **Expected Result:** Whether the tutor should accept or correct the student's final statement.

- **Math Level:** The educational level of the math problem (Elementary, Algebra, Geometry, Trigonometry, Calculus).

- **Conversation Data:** A list of conversation entries, with roles specified as student or tutor.

- **Test ID:** Unique identifier for the test case.

5 Dataset Statistics

This benchmark dataset comprises 188 conversations, with the breakdown shown in Tables 1, 2, and 3.

Math Level Distribution	**Number of Conversations**
Elementary	52
Algebra	45
Geometry	26
Trigonometry	30
Calculus	35

Table 1: Distribution of Math Levels in the Dataset

Expected Result Distribution	**Number of Instances**
Answer Accepted	106
Answer Not Accepted	82

Table 2: Distribution of Expected Outcomes

Conversation Length Statistics	**Value**
Average conversation length	10.77 entries
Maximum conversation length	43 entries
Minimum conversation length	3 entries

Table 3: Conversation Length Statistics

6 Anonymization Process

Creating a diverse, realistic dataset of this type is highly difficult. The problem space that we want to evaluate is very large. The only way to create a dataset that is truly representative of how conversations unfold between a student and a tutor is to start with actual conversations. In order to address privacy concerns regarding use of personal data, each data entry is anonymized and modified to create representative data that cannot be associated with a specific individual. This process includes manual review to ensure that no personal identifiers are included in the conversation, modification of the conversation to alter the phrasing of the question, without changing any of the math or any of the mathematical terms in the conversation, and a final manual review to ensure the math has stayed consistent and that the entry does not include any personally identifying information. Each entry is assigned a unique, non-personal identifier of the conversation.

The dataset entry is then labeled with the expected output from the LLM. The labels used are:

- **Answer Accepted:** The LLM should not indicate the answer is wrong. This can take the form of telling the student good job, or just moving on to the next part of the conversation.
- **Answer Not Accepted:** The LLM should notice that the student made a mistake in the previous message. This typically takes the form of pointing out the error and guiding the user towards the right answer.

7 Evaluation Scenarios

The conversations are structured in a binary format where the LLM either identifies the student's claim as correct or incorrect. This setup was chosen to simplify the evaluation of the LLM's response. Tutoring scenarios can be complex, subtle, and nuanced, with significant variation across different interactions. By structuring the dataset this way, we provide a straightforward, binary method for evaluating responses. This approach not only simplifies the evaluation process but also potentially facilitates better automation.

8 Metric Results

The dataset was employed to assess using several models. To ensure impartial scoring, an LLM-based evaluation was employed. The prompt shown in Listing 1 was used for the evaluation, with the following model settings:

Model Settings:

- Model: `gpt-4-turbo-2024-04-09`
- Temperature: `0.0`

It is important to note that while this prompt and its settings provide a consistent scoring method, it is not perfectly accurate. A sampling indicated that it was approximately 90% accurate.

Listing 1: Tutor Evaluation Prompt

```
This text is a tutor replying to a student's work:

{full_conversation}

Based on the final response
=====
{airesponse}
=====
did the tutor say or imply that the student made a
   mistake when the student said:

{student_response}

Answer with 'yes' or 'no'
```

A note on the process: OpenAI models display variability in results, even with a temperature setting of zero. Consequently, each test item was run ten times for OpenAI models to account for this variation. In contrast, Anthropic models and Phi returned consistent results and were therefore run only once. All temperatures were set to zero.

A note on the Phi model: A fine-tuned version of Phi-3-medium was used in this analysis, it's in private preview and hasn't integrated content data from Khan Academy yet. Phi-3 is Microsoft's family of small language models (SLMs).

Table 4 presents the overall results. A consistent pattern emerges, indicating that LLMs perform better in scenarios where the student is correct than in situations requiring correction of the student's mistakes. Proper prompting can potentially mitigate this skew, although specific prompting techniques are beyond the scope of this paper.

Model	**Student Incorrect**	**Student Correct**	**Total**
gpt-4o-2024-05-13	68.5	85.8	78.3
Fine-tuned Phi-3-medium	72.0	81.1	77.1
gpt-4-turbo-2024-04-09	62.6	86.6	76.1
claude-3-opus-20240229	56.1	89.6	75.0
gpt-4-0613	54.9	84.8	71.8
claude-3-5-sonnet-20240620	50.0	88.7	71.8
gpt-3.5-turbo-0125	41.1	87.7	67.4
claude-2.1	31.7	91.5	65.4

Table 4: Model Overall Performance (pct)

Table 5 details per-subject level results. While GPT 4o scores the highest score overall, it is not universally the best across all domains. GPT 4 Turbo significantly out performs GPT 4o in calculus. GPT 4 outperformed in trigonometry. The fine tuned Phi 3 outperforms in calculus and trigonometry.

Model	Ele.	Algebra	Geometry	Trig	Calc	Total
gpt-4o-2024-05-13	85.2	83.1	73.8	75.0	68.0	78.3
Fine-tuned Phi-3-medium	80.8	71.1	73.1	80.0	80.0	77.1
gpt-4-turbo-2024-04-09	79.6	77.8	70.8	70.3	77.7	76.1
claude-3-opus-20240229	82.7	75.6	69.2	70.0	71.4	75.0
gpt-4-0613	73.7	73.6	68.5	80.3	61.7	71.8
claude-3-5-sonnet-20240620	76.9	68.9	76.9	70.0	65.7	71.8
gpt-3.5-turbo-0125	74.6	73.1	60.8	67.7	54.0	67.4
claude-2.1	80.8	64.4	65.4	56.7	51.4	65.4

Table 5: Model Subject Level Performance (pct)

9 Limitations

We acknowledge that this dataset only evaluates one aspect of tutoring behavior, the evaluation of student responses. It does not in any way evaluate whether the model returns good tutor moves, for example. There is more work to be done to evaluate the many other aspects of tutoring.

10 Acknowledgments

We would like to express our sincere gratitude to everyone who contributed to the creation of this dataset. Your efforts in gathering, processing, and validating the data were invaluable. Special thanks to Charlie Auen, Karen Shapiro, Nick Kokkinis, and Victoria Cheng for their dedicated support and effort for this project. Your contributions have been instrumental in making this work possible. This work was made possible through the support of the Learning Engineering Virtual Institute (LEVI), a sponsored project of the Walton Family Foundation and Rockefeller Philanthropy Advisors.

11 Conclusion

Using LLMs for math tutoring presents unique challenges that go beyond solving math problems. The proposed benchmark dataset aims to provide a robust framework for evaluating and improving accuracy of LLMs when evaluating math in tutoring contexts. By addressing this specific challenges, we can develop AI tutors that are not only supportive but also mathematically accurate in real-time tutoring scenarios, thereby enhancing the overall effectiveness of LLMs in assisting students with their math difficulties.

References

[1] "ChatGPT has entered the classroom: how LLMs could transform education." Nature, 623, 474-477 (2023). doi: 10.1038/d41586-023-03507-3.

[2] Hendrycks, Dan, et al. "Measuring Mathematical Problem Solving With the MATH Dataset." arXiv preprint arXiv:2103.03874, 2021.

[3] Cobbe, Karl, et al. "Training Verifiers to Solve Math Word Problems." arXiv preprint arXiv:2110.14168, 2021.

[4] Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

[5] Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495-522.

[6] Vygotsky, L.S. "Mind in Society: The Development of Higher Psychological Processes." Harvard University Press, 1978.

[7] Brown, Tom B., et al. "Language Models are Few-Shot Learners." arXiv preprint arXiv:2005.14165, 2020.

[8] Ziegler, Daniel M., et al. "Fine-Tuning Language Models from Human Preferences." arXiv preprint arXiv:1909.08593, 2019.