**Technical Note: Current Ability and Possibilities of LLMs in Education**

**Author Names and Affiliations:**

Aarush Shrikanth¹,

¹Los Altos High School, 827 Orchid Place, Los Altos, CA 94024, USA, aarush.shrikanth@gmail.com

**Corresponding Author:**

Aarush Shrikanth

aarush.shrikanth@gmail.com, +1 650 447-9901, 827 Orchid Place, Los Altos, CA 94024, USA  
  
**Abstract**

This paper explores the possibility of using Large Language Models (LLMs) in math education. Specifically, it explores the current ability of widely used models to guide students towards the correct solution in elementary level word problems. It also explores how these models could be improved for this goal through methods like hidden prompts (also known as system prompts) (4) and fine-tuning. These findings are presented through both qualitative and quantitative analysis, using semantic similarity for the quantitative analysis.. Finally, we discuss some of the limitations of LLMs in this use case, that are still not handled by fine-tuning and prompting. The data overall demonstrates the most success with prompting and some success with fine-tuning, although the model's responses were too inconsistent to be deployed on its own at this time.  
  
**Keywords:**

Large Language Models, Math Education, Prompting, Fine-tuning, Semantic Similarity  
  
**1. Introduction**

The ability of artificial intelligence and neural networks to mimic human use of language (5) has revolutionized many parts of the world, making tasks that once seemed impossible to automate accessible. One such area is education..

This paper aims to explore the ability of LLMs to give feedback and guide students toward the correct answer, in the context of elementary level math education. Methods including fine-tuning and prompting are used to improve the models current capabilities, bringing the LLMs closer to simulating real teacher-student interaction.

Through these methods, the paper explores the extent to which current LLMs are able to help students arrive at the correct answer, and how changes to the models subsequently change their ability to achieve this task.

**2. Methods**

**2.1 Fine-tuning**

The data set used for training and testing data is a dialogue tutoring data set, composed of math word problems along with a student-teacher conversation in which the teacher guided the student to the correct answer (1).  
  
To form conversations, the fine-tuned model was given a series of messages that contained a system prompt with the question the conversation was based on, and then subsequent messages between a student and the teacher until the student obtained the correct answer. Through this method, the model was trained with the goal of improving its ability to ask guiding questions that help students find the answer on their own.  
  
**2.2 Prompting**  
The prompt chosen was:

“You are a teacher supposed to help guide students towards the right answer. The problem the student is working on is {problem}. Do not tell the student exactly where they went wrong, instead, guide them towards recognizing their mistake so they can fix it themselves. Ask them questions about the way they went through the problem so that they recognize their mistakes, but keep your response concise and to the point”  
  
This prompt was chosen to lead the model towards asking guiding questions, helping students reach the answer independently without explicit solutions being given.

**2.3 Evaluation**

Performance was measured using two distinct methods: **Semantic similarity** (via Cohere’s "embed-english-v3.0" embeddings) comparing model responses to ideal teacher responses, and **Qualitative scoring** assessing guiding effectiveness.

**3. Discussion and Results**  
  
**3.1 Example Case Study**  
Example 7 in the data set (2) contains the following problem, with the correct answer of 7: "There are 4 snails in one aquarium and 32 snails in another aquarium. The difference between the number of snails in the two aquariums is twice the amount of fish in both aquariums. If both aquariums have the same number of fish in them, how many fish are there in each aquarium?".

The students answer was mostly correct in the setup of the system, and they arrived at 14 fish for their solution, forgetting that 14 is the total amount of fish, and the question is asking for the amount in each tank. The given teacher's answer points out this exact mistake, stating that the student had forgotten to consider the questions' nature. The student continued the conversation without correctly fixing their answer.

In the next turn, the unprompted and non-fine-tuned model simply states

“The student is correct in their reasoning. To recap: 1. The number of snails in the first aquarium is 4. 2. The number of snails in the second aquarium is 32. 3. The difference in the number of snails between the two aquariums is \( 32 - 4 = 28 \). The problem states that this difference is twice the number of fish in both aquariums. If we let \( f \) represent the number of fish in each aquarium, then the total number of fish in both aquariums would be \( 2f \). Setting up the equation: \[ 28 = 2f \] To find \( f \), we divide both sides by 2: \[ f = \frac{28}{2} = 14 \] Thus, each aquarium has: \*\*14 fish.\*\*”.

The model just restates the students' reasoning, and was incorrect as it claimed that the students' final answer of 14 fish was correct.

On the other hand, the next model, promoted but non-fine-tuned, pointed out the students success in determining the total number of fish while still pointing out their mistake, saying that the student

“found a connection between the difference and the number of fish! However, have you carefully considered how the problem defines the relationship between the difference and the number of fish? What exactly does it say about the amount of fish in both aquariums?"

In this way, the model took advantage of a guiding question to lead the student toward determining the answer properly.

The unprompted but fine-tuned model also pointed out their success in determining the difference between the number of snails and mentioned that the student was missing a step, saying

“Okay, so you're correct, the difference is 28. And you're on the right track with saying that twice the number of fish is 28. But you forgot to add one step to the equation. What is the opposite operation of division?”

However, its guiding question, “What is the opposite operation of division?”, was much more vague and difficult to interpret in the context of the problem.

Finally, the prompted and fine-tuned model responded with an incomplete answer:

"Yes, but if I told you there are two aquariums and they have the same number of fish that would suggest".

This response leaves room for the student to finish the sentence, which in theory could work but is less clear and is also hard to understand while working through the problem itself.  
  
**3.2 Prompting Effectiveness**

One of the key problems with the models prior to any changes was its tendency to give up the answer to the student immediately. Prompting was able to solve this problem by making sure the model was able to keep its response focused on guiding the student. In addition, it also led to the model pointing out the issues with the students' past work, which could have the effect of making students waste less time recognizing the mistake and instead being able to continue with their work.

**3.2 Fine Tuning**  
Fine Tuning on the other hand had a different effect, the teacher responses used to finetune the model composed of just singular questions asking the student for their strategy, and therefore the fine tuned models just asked the student to walk through their answer/explain their reasoning for their answer. This was not super effective, as it focused too much on conciseness without helping the student get the correct answer.

For this reason, prompting was used in conjunction with fine tuning in one of the tests. Models that had both been finetuned and given a prompt were still concise, but in some cases asked more specific questions that focused on where the student went wrong, similar to one part of the promoted and not finetuned models answers.

**3.4 Evaluating effectiveness through embeddings**

The semantic similarity calculated by embeddings in Figure 1 represents how close the models answers are to the teachers answers, and by dividing it by the normal model unprompted, it demonstrates how similarity changes from the normal model through certain methods. The normal model prompted in this case is clearly the only instance where the model improved in similarity to the teacher's response with respect to the original, normal and unprompted, model.

**3.4 Qualitative Analysis**

The LLMs responses were also scored to gauge the normal and fine tuned model's, both with prompting, effectiveness. The scoring corroborated the results found from the embeddings, demonstrating that just prompting was the most successful, while fine tuning and fine tuning with prompting were less successful (3).

**3.4 Practical Implications**

Understanding the differences in these two methods is crucial to finding out the correct fields to apply them in. Fine-tuning models intended for education in more complex and specific fields, such as medicine or law, could be useful (6, 7) because it could add both additional knowledge and improve the structure of the LLMs responses relative to the topic at hand. On the other hand, using prompting on its own is much simpler, and could be an important factor to making this form of teaching much more accessible for many, reducing expensive and possibly unnecessary resource usage.

Although the data demonstrates that both of these methods were successful in improving the models, the inconsistency of the models' success makes it clear that the LLMs evaluated cannot be left alone to teach students. Therefore, if these methods were put into practice, further advancements or fail safes based on human interactions would be crucial to guaranteeing successful and consistent results.  
  
**4. Conclusion**

As a result of the advancement of Artificial Intelligence and Large Language Models in recent years, it is crucial to take full advantage of these new opportunities in education, enabling students to achieve much more than ever before (8). This paper explores the current capabilities of these models, and the extent to which they could be improved through specific methods such as fine-tuning and prompting. The data demonstrated that current models partially lack the ability to help students reach their own conclusions, and that to successfully improve students' ability, they need to improve their ability to push students in the right directions. Fine-tuning and prompting helped the LLMs improve their effectiveness in helping students reach the correct answer on their own, without outside intervention. Although both methods had effects on the LLMs, prompting ended up being more successful because the prompted models asked guiding questions that would have led the student to the correct answer using the student's own ability.

In terms of future work, we’d like to point to recent work in “chain of thought” reasoning (9). Chain of thought prompting showed promising results and this was followed by using reinforcement learning to reward the model for developing complex thought processes. This has led to complex reasoning models like openai-o1 and deepseek-r1. We believe a similar technique can be applied to this process by rewarding the development of guiding questions during models training, so that they can better be applied to education and represents an interesting area of future investigation.

### **Acknowledgements**

I would like to thank my mentor, Dr. Daniel Raggi , for advising me on both my writing and experimentation processes.

### **References**

1. Macina J, Daheim N, Chowdhury S, Sinha T, Kapur M, Gurevych I, Sachan M. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics, pp. 5602–5621, 2023.
2. Shrikanth A. Dataset. 2025. Available from:https://github.com/EmmetAVS/AI\_ED. Accessed: 2025-6-19.
3. Shrikanth A. Qualitative Scoring CSV. 2025. Available from: https://github.com/EmmetAVS/AI\_ED/blob/main/qualitative\_scoring.csv. Accessed: 2025-6-19.
4. Ramlochan S. System Prompts in Large Language Models. Available from:<https://promptengineering.org/system-prompts-in-large-language-models/>.
5. Zheng M, Pei J, Jurgens D. Is "a helpful assistant" the best role for large language models? A systematic evaluation of social roles in system prompts. *arXiv preprint*, arXiv:2311.10054, 2023.
6. Holmes W, Tuomi I. State of the art and practice in AI in education. *European Journal of Education Research Development and Policy*, 57(4):542–570, 2022. doi:10.1111/ejed.12533.
7. Selwyn N. The future of AI and education: Some cautionary notes. *European Journal of Education Research Development and Policy*, 57(4):620–631, 2022. doi:10.1111/ejed.12532.
8. Bettayeb AM, Talib M, Altayasinah SZ, Dakalbab F. Exploring the impact of ChatGPT: conversational AI in education. *Frontiers in Education*, 9, 2024. doi:10.3389/feduc.2024.1379796.
9. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, Denny Zhou: Chain of Thought Prompting Elicits Reasoning in Large Language Models. CoRR abs/2201.11903 (2022)