February 15th, 2024

Dear Reviewers,

I am grateful for the opportunity to respond to the insightful feedback provided on the paper, "Zero-Shot Mathematical Problem Solving with Large Language Models via Multi-Agent Conversation Programming." Your constructive comments are instrumental in underscoring the broader impact and potential applications of the work.

Enhancing AI Tutors for Education — First I would like to address the valid concern regarding how this work directly contributes to applied education, particularly since mathematical problem solving is often used as a benchmark for recent LLM advancements. Our main motivation for pursuing this line of research has been to expand the accessibility of math education through the creation of an AI tutor. By introducing a novel multi-agent conversation programming approach, we address a critical flaw in many current AI educational tools: that often times they arrive at incorrect solutions.

Furthermore, I believe there are many advantages to our approach when considering the potential impact on educational tool design, in particular adaptive tutors. We not only arrive at the correct numerical solution more often, but in the process also generate step-by-step dialogue (not just step-by-step solutions) with pedagogical value. In practice, these extended dialogues are valuable to guide an AI tutor towards a more socratic style rather than a solver agent. One can imagine using our Aurek conversation programming strategy to support an "infinitely patient AI tutor" agent having a dialog directly with the student, clarifying questions, providing rationales and overall engaging the student far more than just a solver.

Comparison with other published benchmarks results — Although several other papers have reported superior headline numbers, I would like to emphasize that these are often achieved through best-of-many or few-shot approaches. Some of these solve the same problem as many as a thousand times which is expensive and not aligned with providing accessible educational tools, like individual tutoring, to a mass audience. Approaches that involve retraining models are also expensive and involve access to model weights. While we reference these strategies in our paper as having better accuracy, our approach was done as best-of-one, zero-shot, and is implementable without access to model weights. There is some additional overhead with multi-agent conversation programming, but we still believe it is less expensive overall. Finally, when the same sources we reference used a best-of-one inference their results were no better than ours. This suggests to us, that with further innovation on our approach we could see increased accuracy over time while still employing a zero-shot strategy.

We hope these clarifications and additional details address the concerns and suggestions raised by the reviewers. Our research lays the groundwork for future investigations into the application of AI in educational settings, particularly in the realm of mathematics education, where accuracy and reliability are crucial.

Sincerely,

Vivian Keating
GotIt! Education
vivian@gotitapp.co