ENHANCING MATHEMATICAL ASSESSMENT ACCURACY THROUGH MULTI-AGENT SYSTEMS: A COLLABORATIVE APPROACH TO EVALUATING STUDENT SOLUTIONS *

Hon Kit Long
The University of Hong Kong
Hong Kong
u3608018@connect.hku.hk

Au Chi Kin
Hong Kong University of Science and Technology
Hong Kong
ckauac@connect.ust.hk

ABSTRACT

This paper explores the enhancement of mathematical assessment accuracy through the use of multi-agent systems (MAS) combined with large language models (LLMs). Recent advancements in LLMs, particularly techniques like chain-of-thought reasoning and the multiple attempts approach, have significantly improved their problem-solving capabilities. We propose a system named MATH-marker, which utilizes interpretable agents to provide calculation references for grading student solutions. Our approach aims to evaluate student work from multiple perspectives, leveraging the strengths of MAS to enhance accuracy and reliability. Our experiments reveal a strong correlation between the model's confidence levels and the accuracy of its evaluations, suggesting that higher confidence can serve as a reliable indicator of correctness. Despite promising results, we acknowledge limitations in sample size and the potential impacts of zero-shot prompting on performance. Future research should focus on larger datasets and the fine-tuning of models to further enhance the effectiveness of mathematical assessments in educational contexts.

***K*eywords** Multi-Agent Systems · Large Language Models (LLMs) · Education

1 Introduction

Recent studies have highlighted the potential of using large language models (LLMs) in various applications, including mathematics [1]. Techniques such as chain-of-thought reasoning and the divide-and-conquer method have significantly enhanced the ability of LLMs to solve mathematical problems [2,3]. This paper explores the use of interpretable agents as a means to provide calculation references for LLMs in grading student work. By leveraging the divide-and-conquer approach, we aim to further improve the performance of LLMs in educational settings [3,4].

* *Citation*: **ENHANCING MATHEMATICAL ASSESSMENT ACCURACY THROUGH MULTI-AGENT SYSTEMS: A COLLABORATIVE APPROACH TO EVALUATING STUDENT SOLUTIONS , Hon** & **Au**

2 Previous Work

2.1 LLM in Eduation

Large language models (LLMs) have demonstrated significantly improved performance across various metrics, particularly with recent advancements. [5–8] Research indicates numerous opportunities to integrate LLMs into educational contexts [1, 9], including applications such as self-learning assistants, grading tools, and writing support. Certain models have been specifically developed to interact more effectively with students when addressing mathematical questions [10]. Overall, LLMs appear to be more relevant and applicable in education today than in the past.

2.2 Multi agent

Recently, the use of multi-agent systems that incorporate large language models (LLMs) has gained traction, with applications in frameworks, problem-solving, world simulation, and more [11]. Various communication methods for these agents have been developed, including centralized, layered, decentralized, and shared message pool approaches [11]. Studies have shown that increasing the number of agents can enhance accuracy across various domains, including mathematics and chess [12]. This provides a critical foundation for developing a solution checker that utilizes multi-agent systems.
Some multi-agent systems are specifically designed for educational purposes, dividing tasks into multiple sections that encompass areas such as logic and reasoning [13]. These systems enhance the learning experience for students, as indicated by recent studies [13]. Additionally, algorithms have been developed to enable agents to learn, allowing them to function as both students and teachers simultaneously [14]. Overall, multi-agent systems facilitate the interaction of LLMs with various environments in a more efficient, flexible, and effective manner.

2.3 Prompting Method

There exist numerous prompting methods, each contributing to an overall enhancement in accuracy [2, 3, 15, 16]. These techniques facilitate a more precise interaction with large language models (LLMs), thereby enabling us to bolster their performance. Additionally, few-shot prompting has demonstrated significant improvements in LLM efficacy [17]. Moreover, prompting allows us to define the role of the LLM, establishing a foundation for the subsequent development of our agent.

2.4 Mathematics and LLM

Research has shown that the performance of basic arithmetic operations is not satisfactory for large language models (LLM) [18]. However, some studies demonstrate that employing a divide and conquer method for mathematical tasks has achieved state-of-the-art results [3]. Additionally, research indicates that utilizing Lean as an agent for proving theorems with Lean can enhance accuracy [19].

2.5 Research Gap

Currently, there is no model that combines a multi-agent approach with divide and conquer methods in Python for solving marking solutions. We aim to explore this approach to evaluate whether the overall performance is acceptable. We call our system MATH-marker.

3 Structure

The MATH-marker is composed of four parts. First, we will employ DeepSeek-V3, known for its high performance in mathematics [7]. We utilize multi-agent methods to establish connections between the answers and questions, allowing us to approach each question from five different perspectives and identify the dominant response.

In this experiment, we continue to use zero-shot prompting. Our goal is to develop a more general multi-agent structure rather than a fixed-domain approach. Consequently, we focus on zero-shot prompting as a key method.

3.1 Experiment

We employ a method that requires the LLM to evaluate the correctness of an answer based on several aspects: contrapositive reasoning, numerical analysis, axiomatic foundations, contradictions, and logical reasoning. The dominant evaluation among these aspects determines the final answer. Our data source for this method is MATH80 [20].

3.2 Result

The results from our experiment reveal significant insights into the performance of our model. Out of a total of 80 evaluations, 47 responses were classified with a confidence level of 1, indicating that these answers are deemed to be 100% correct. This high confidence suggests that the model can reliably provide accurate answers in many cases. However, there were 2 instances where the answers were incorrect, with confidence levels of 0.66 and 0.58, respectively. These lower confidence scores highlight the model's uncertainty in those particular cases, suggesting that even when the model is less confident, it is still capable of identifying potential inaccuracies.

Furthermore, the remaining responses were correct, and an interesting trend emerged: higher confidence levels generally corresponded to more accurate answers. This observation implies that the model's confidence can serve as a useful indicator of answer validity, allowing us to prioritize responses based on their confidence ratings. In summary, while the model demonstrates strong performance overall, the correlation between confidence levels and accuracy underscores the importance of examining confidence as a critical factor in assessing the model's outputs.

For a detailed overview, refer to the table below:

Evaluation	Count	Confidence Level
Correct (Confidence 1)	47	100%
Incorrect	2	0.66, 0.58
Correct (Other)	31	Varies

Table 1: Summary of Evaluation Results

For more information, please visit our GitHub repository:

`https://github.com/JasonAlbertEinstien/MATH-MARKER/tree/main`

4 Limitation

While our study demonstrates promising results, several limitations must be acknowledged. First, the reliance on zero-shot prompting may restrict the model's performance in scenarios where context or specialized knowledge is essential. Without fine-tuning on specific datasets, the model may struggle to provide accurate responses in more complex mathematical problems.

Additionally, the evaluation process involved a relatively small sample size of 80 questions, which may not sufficiently represent the broader range of mathematical concepts. This limited dataset could affect the generalizability of our findings. Future studies should consider larger, more diverse datasets to validate the model's performance across various mathematical domains.

Moreover, the confidence levels assigned to the answers, while informative, can be influenced by the model's inherent biases and limitations in understanding nuanced queries. As such, reliance solely on confidence scores without further validation could lead to incorrect conclusions about the accuracy of certain answers.

Lastly, the multi-agent approach, while effective in exploring different perspectives, may introduce variability in responses that complicates the interpretation of results. It is essential to further investigate how different agents contribute to overall accuracy and whether their interactions enhance or detract from the model's performance.

5 Conclusion

In this study, we explored the potential of multi-agent systems combined with large language models (LLMs) to enhance the accuracy of mathematical assessments. Our findings indicate that the MATH-marker system demonstrates strong performance, particularly when employing zero-shot prompting. The results revealed a significant correlation between the model's confidence levels and the accuracy of its answers, suggesting that higher confidence can serve as a reliable indicator of correctness.

Despite the promising outcomes, our study is not without limitations. The reliance on a small sample size and zero-shot prompting may impact the generalizability of our findings. Future work should focus on larger datasets and the fine-tuning of models to address these limitations. Moreover, exploring the interactions between different agents within the multi-agent framework could further enhance the robustness of our approach.

Overall, our research contributes to the growing body of literature on the application of LLMs in educational contexts and opens avenues for future exploration in improving the assessment of student solutions in mathematics. We believe that continued advancements in this field can lead to more effective educational tools that support both educators and learners.

References

[1] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.12345*, 2024. Submitted on 26 Mar 2024 (v1), last revised 1 Apr 2024 (this version, v2).

[2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. [Submitted on 28 Jan 2022 (v1), last revised 10 Jan 2023 (this version, v6)].

[3] Hon Kit Long, Au Chi Kin Kinson, Liu Peter Hong, and Cho Chung Hei. Enhancing mathematical problem solving with large language models: A divide and conquer approach. `https://github.com/JasonAlbertEinstien/DaC-LLM`, 2025. Licensed under *Creative Commons Attribution 4.0 International License*.

[4] Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. Mario: Math reasoning with code interpreter output – a reproducible pipeline. 2024. [Submitted on 16 Jan 2024 (v1), last revised 21 Feb 2024 (this version, v3)].

[5] Anthropic. Claude 3.5 sonnet. GitHub, 2024.

[6] Meta AI. Llama 3.3 model card. GitHub, 2024.

[7] DeepSeek AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[8] OpenAI Research Team. Do not think that much for 2+3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

[9] Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. Large language models in education: Vision and opportunities. 2023. [Submitted on 22 Nov 2023].

[10] Murong Yue, Wijdane Mifdal, Yixuan Zhang, Jennifer Suh, and Ziyu Yao. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. 2024. [Submitted on 10 Apr 2024].

[11] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. 2024. [Submitted on 21 Jan 2024 (v1), last revised 19 Apr 2024 (this version, v2)].

[12] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. 2024. [Submitted on 3 Feb 2024 (v1), last revised 11 Oct 2024 (this version, v2)].

[13] Yuan-Hao Jiang, Ruijia Li, Yizhou Zhou, Changyong Qi, and Hanglei Hu. Ai agent for education: von neumann multi-agent system framework. 2024. arXiv:2501.00083v1 [cs.MA], License: arXiv.org perpetual non-exclusive license, submitted on 30 Dec 2024.

[14] Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P. How. Learning to teach in cooperative multiagent reinforcement learning. 2018. [Submitted on 20 May 2018 (v1), last revised 31 Aug 2018 (this version, v4)].

[15] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. 2024. [Submitted on 18 Aug 2023 (v1), last revised 6 Feb 2024 (this version, v4)].

[16] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, and Karthik Narasimhan Yuan Cao. Tree of thoughts: Deliberate problem solving with large language models. 2023. Submitted on 17 May 2023 (v1), last revised 3 Dec 2023 (this version, v2).

[17] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020. [Submitted on 28 May 2020 (v1), last revised 22 Jul 2020 (this version, v4)].

[18] Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. Adapting large language models for education: Foundational capabilities, potentials, and challenges. 2023. [Submitted on 27 Dec 2023 (v1), last revised 26 Apr 2024 (this version, v3)].

[19] Peiyang Song, Kaiyu Yang, and Anima Anandkumar. Towards large language models as copilots for theorem proving in lean. *arXiv preprint arXiv:2404.12534*, 2024.

[20] Au Chi Kin Kinson. Math80 dataset, 2025.