Qeios

Peer Review

Review of: "Diversity of Thought Elicits Stronger Reasoning Capabilities in Multi-Agent Debate Frameworks"

Shuang Liang[1]

1. University of Electronic Science and Technology of China, China

This paper explores how diverse thoughts of LLMs enhance the reasoning abilities within a multi-agent debate framework. Based on the framework by Du et al.[1], this study experimentally verifies the impact of different model scales and diversity on the performance of mathematical reasoning tasks. The main contribution is demonstrating that a diverse set of models can significantly improve reasoning accuracy in debates, even surpassing a single high-performance model like GPT-4. Additionally, this paper finds that even medium-sized models can benefit from multi-agent debates, offering potential applications in resource-constrained environments.

Strengths:

1. **Empirical Support**: The paper provides strong evidence for the effectiveness of its approach through experiments on various mathematical reasoning benchmarks.
2. **Applicability Across Model Scales**: The study shows that the multi-agent debate framework is applicable not only to large-scale models but also to medium and small models, increasing its general applicability.

However, there are several concerns that need to be addressed:

Main Concerns:

1. **Lack of Novelty**: This paper is a straightforward extension of Du et al.'s framework[1]. Can we regard it as a very similar framework by the following setting: replacing the model in Du et al.'s framework with those mentioned by the authors (e.g., Gemini-Pro, Mixtral 7B×8, and PaLM 2-M)? How does its performance differ from the proposed method?

 doi.org/10.32388/MZ1KBG

2. **Missing Baselines**: There is a lack of comparison with more agent-based baselines. The authors seem to focus more on controlled experiments within their method rather than comparing with other advanced methods.
3. **Unclear Definition and Measurement of Model Diversity**: Although the paper emphasizes the importance of model diversity, it lacks a deep discussion on defining and distinguishing "diverse" models. For instance, it is unclear why the closed-source Gemini-Pro and the open-source Mixtral 7B×8 can be considered a pair for debate. This leads to a lack of standardization in model selection (can it be assumed that the authors chose models favorable to their framework?).
4. **Complexity and Resource Consumption**: Implementing the multi-agent debate framework is relatively complex and may require significant computational resources. This also relates to the lack of baselines; if a single GPT-4 call can achieve reasonable reasoning, does the improvement from the authors' method justify the time and space costs?

Minor Concerns:

1. **Uncertainty in Generalization**: This paper mainly focuses on solving mathematical problems, lacking discussion on the framework's generalization to other types of reasoning tasks.
2. **Collaboration Among Models**: Does collaboration among models lead to correct agents being influenced by incorrect ones? The exploration should not be limited to correcting errors but also consider whether correct results are induced to become incorrect.

References:

[1] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:230514325.

Declarations

Potential competing interests: No potential competing interests to declare.