Majority Errors in Multi-Agent Debate: Analysis and Framework Design

Meitong Liu
meitong4@illinois.edu

Jieyi Zhao
jieyi3@illinois.edu

Wangjia Zhan
wangjia2@illinois.edu

Maojie Xu
maojiex2@illinois.edu

Ian Jiang
jisheng3@illinois.edu

Abstract

Multi-agent debate (MAD) has been shown to improve the reasoning abilities of large language models (LLMs) by enabling multiple agents to exchange responses and reach consensus. However, a more challenging setting has been rarely examined: when most agents initially produce incorrect answers, which we refer to as majority error. This project studies how MAD behaves in this challenging scenario, evaluates its performance as the number of agents and debate rounds increases, and reveals that the gains arise from additional sampling instead of the debate process. We further explore improvements such as adding confidence scores from an external critic model and introducing specialized roles that encourage diverse reasoning styles. These additions make the debate more stable and lead to better outcomes on the majority-error tasks. Finally, we link the observed limitations of MAD compared to repeated sampling to recent theoretical explanations, emphasizing and verifying that MAD's potential may come from high-quality supervision. Code is released in `https://github.com/nilgeoutim/CS546_MajorityErrorDebate`.

1 Introduction

Multi-agent Debate (MAD) frameworks have emerged as a promising approach for improving the reasoning abilities of Large Language Models (LLMs) (Du et al., 2023; Chan et al., 2023; Khan et al., 2024). In these systems, multiple LLM agents engage in iterative discussions to refine their initial answers and converge on a solution through a majority vote. MAD has been reported to not only boost accuracy on complex reasoning problems and factual QAs (Du et al., 2023), but also generalize to enhance performance in related tasks, such as translation (Liang et al., 2024) and negotiation (Fu et al., 2023), and assist in model self-improvement (Subramaniam et al., 2025).

However, a critical situation has been less examined: how effective is MAD when the majority of agents initially produce incorrect responses and only a few hold the right point? This would happen when the system is facing particularly challenging problems. Do the correct tend to conform to the opposite majority, leading to a "wisdom of crowds" failure where the entire system converges on an incorrect consensus? Previous studies reported improved accuracy of MAD systems over majority voting without debate (Du et al., 2023), indicating that a system where the incorrect outweigh the correct can still benefit from collaborative reasoning. Nevertheless, these gains are often inconsistent across models and datasets, and marginal compared to the remaining percentage that majority voting fails to solve (Wynn et al., 2025). In addition, the influencing factors and underlying drivers, such as the total number of times a model is invoked, i.e., the effective sample size, remain unclear.

Taken together, these observations point to a gap in our current understanding of MAD. Specifically, when agents face problems on which most initial responses are incorrect, it remains unclear how effective MAD is under such majority-error conditions, how its performance scales as we vary the number of agents or debate rounds, and what interventions might help strengthen the system against erroneous consensus formation. Clarifying these questions is essential for evaluating the robustness of MAD on genuinely difficult tasks and for guiding the development of more reliable debate-based reasoning frameworks.

In this project, we focus on the setting where the initial majority of agents is incorrect. We construct a subset of GSM8K containing problems for which a three-agent majority vote fails, and use it as a testbed to conduct a detailed evaluation of how MAD behaves under majority-error conditions. Our analysis shows that naïve MAD exhibits strong fluctuations across debate rounds, that correct agents flip to incorrect answers at nearly the same rate as the reverse, and that accuracy gains are largely explained by increased sampling rather than the debate process itself. Under the same computation budget, i.e., the same effective sample size, MAD is outperformed by majority voting among independently sampled responses.

Beyond analysis, we investigate simple modifications to stabilize debate. We incorporate confidence scores from a separate critic model and introduce specialized reasoning roles to encourage diverse perspectives among agents. Both additions effectively provide a more stable debate on the difficult problems, yielding consistent improvements across debate rounds and increasing the final accuracy.

Finally, we refer to a recent theoretical analysis of MAD (Choi et al., 2025) to explain its observed disadvantage relative to majority voting—debate alone cannot increase agents' belief in the correct answer and thus cannot lead the process to converge to a right consensus, while increasing the number of agents in majority voting exponentially increases the probability of obtaining a correct answer. Accordingly, we propose and verify through experiments that the potential of MAD must come from the system's ability to grow more confident in the correct direction, potentially through high-quality external supervision, such as minimally incorporating expert feedback.

Overall, our project provides a deeper understanding of why MAD struggles in majority-error scenarios and demonstrates practical directions for making debate-based systems more reliable on challenging reasoning tasks.

2 Related Work

Existing work related to whether MAD can improve performance can be broadly categorized based on their positive or negative views.

Debate as a performance enhancer. Multi-agent debate is shown to improve the performance of accuracy, reasoning, as well as reliability by enabling agents to be exposed to diverse solution paths, reconsider and resolve errors to reach a more reliable response than single agents. There is existing work supporting this claim, including: task-solving problems like math and commonsense where agents iteratively critique and refine each other's reasoning, structured prompts that encourage different thinking paths, and multi-judge evaluation setups that combine arguments to reduce single-judge noise (Du et al., 2023; Liang et al., 2024; Chan et al., 2023).

Limits and failure modes. Some recent work shows that these benefits mentioned do not always hold. In groups with heterogeneous agents, debate may lower accuracy when weaker agents provide bad arguments that sway stronger ones (Wynn et al., 2025). Furthermore, LLM judges and debaters show style biases, including verbosity, earlier positions, and confidence tone that make persuasive but wrong arguments more influential (Saito et al., 2023; Shi et al., 2025).

Methods for improving debate. A variety of techniques have been proposed to increase debate robustness, including introducing external feedback signals and encouraging better confidence calibration. Recent work shows that explicitly modeling or expressing confidence can improve self-consistency and debate quality (Taubenfeld et al., 2025; Lin and Hooi, 2025). Other research promotes the use of diverse reasoning strategies in multi-agent setups (Liang et al., 2024), and demonstrates that structured roles or curated reasoning traces can strengthen agents' ability to critique and defend arguments effectively. Building on these directions, our work examines the performance of MAD in majority-error scenarios and demonstrates that such lightweight mechanisms—confidence scoring and role specialization—are indeed effective in stabilizing debate and improving performance.

MAD in majority-error settings. Despite growing interest in debate-based methods, relatively little work focuses on how debate behaves when most agents begin with incorrect answers. Existing studies on tasks with general difficulties exhibit diverse findings: some reports MAD outperforms majority voting among the same number of agents, indicating an improvement on the challenging problems (Du et al., 2023), while others report unstable debate processes among heterogeneous agents with correct positions overturned (Wynn et al., 2025). Nevertheless, no prior work has closely examined the dynamics of MAD on purely majority-error questions, which can best reflect its distinction from majority voting. This gap motivates our evaluation of MAD specifically under majority-error conditions, as well as systematic improvements that remain useful in these difficult settings.

3 Proposed Approach

In this section, we elaborate on the two extensions to the standard MAD framework that aim to improve its performance in the majority-error setting, namely external confidence scoring and role specialization. We defer the detailed method for the systematic evaluation of naive MAD to Section 4.

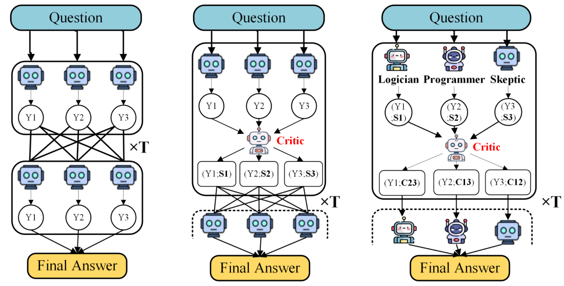
3.1 Confidence Score

One weakness of naive MAD is that all responses are treated equally, regardless of their reasoning quality. To mitigate this, we implement an **external scoring mechanism**. Specifically, after each debate round, agents submit their responses to an external critic, which evaluates and compares logical coherence and mathematical correctness and assigns a scalar confidence score (1–10) along with brief feedback. As shown in Figure 1(b), these scores are fed back into the next debate round and guide agents to update their reasoning in a score-aware manner: high-confidence agents tend to defend and refine their previous answers, while low-confidence agents more aggressively incorporate insights from higher-scored peers.

Beyond influencing belief updates, the confidence signal also enables adaptive control of the debate process. If all agents converge to the same high-scoring answer in the initial round, the problem is deemed easy, and the debate terminates early. Conversely, if all responses receive uniformly low scores, this indicates that the initial sampling likely missed a correct solution, and the round is discarded in favor of resampling. As a result, the confidence score not only improves answer accuracy but also allows computation to be dynamically allocated based on problem difficulty—quickly resolving easy cases while dedicating more debate rounds to harder ones.

While prior work, such as ReConcile (Chen et al., 2024), also incorporates confidence scores, our approach is distinct in that the scores are produced by an explicit external critic. This critic has an oversight view of all agents' responses and performs cross-comparison before assigning scores, enabling more informed and globally consistent evaluations.

Our results in Section 5 show that incorporating critic scores leads to consistent performance gains across rounds, especially when a stronger model (GPT-5.1) is used as the critic.



(a) Multi-Agent Debate (b) MAD w/ confidence score (ours) (c) MAD w/ confidence score & specialization (ours)

Figure 1: Overview of our debate frameworks: (a) naive MAD, (b) debate with confidence score, (c) debate with confidence score and role specialization. Y: output(answer and reasoning), S: confidence score, C: comment and feedback.

3.2 Role Specialization

Another limitation of naive MAD is the homogeneity of agents: identical prompting leads them to follow similar reasoning paths, making the system vulnerable to collective errors. To address this, we introduce role specialization. Through preliminary experiments, we identified that core reasoning roles like *Logician* and *Skeptic* are essential. We also attempted domain-specific personas such as "Financial Analyst" and "Mathematician"; however, these roles often resulted in over-specialization with poor generalization, where agents failed to apply role-specific knowledge flexibly across diverse mathematical contexts. Inspired by prior work on structured reasoning and diverse deliberation, we explore a simple form of role specialization as shown in Figure 1(c).

Consequently, we converged on three complementary:

- **Logician**: Serves as the deductive baseline, utilizing standard Chain-of-Thought (CoT) to break down complex problems sequentially.
- **Programmer**: Solves the problem via executable Python code. Forcing the output into a code format compels the agent to make its logic and specific calculations concrete and precise.
- **Skeptic**: Utilizes *Negative Constraints*. Instead of solving directly, the Skeptic is prompted to first describe plausible but incorrect approaches.

These roles enforce diversity in reasoning structure and reduce the chance that all agents fall into the same incorrect pattern. As shown later, this specialization produces more stable improvements over debate rounds and outperforms the naive MAD baseline on GSM-MajorityError.

Implementation Details We implement these roles via system-level prompting. Specifically, the Programmer leverages the "Program-of-Thought" (PoT) approach, which exposes the logical chain clearly and significantly reduces hallucinations common in natural language generation. Furthermore, the structured code output assists agents in more accurately assessing the *Logic Score* and *Computation Score* during the debate phase, as logical flows and execution steps are explicitly separated. The Skeptic's negative constraints mechanism preemptively blocks common trap answers and forces the model to verify its path against known pitfalls. The specific instructions provided to the agents are as follows:

> **Logician**: *"You are a logical thinker. Solve this problem step-by-step. Break down complex logic into simple, sequential steps."*
>
> **Programmer**: *"You are a Python expert. Write a Python script to solve this math problem. Then, deduce the final answer from your code logic and output it."*
>
> **Skeptic**: *"You are a critical reviewer. Use 'Contrastive Chain-of-Thought' reasoning. Task: 1. First, describe 2 plausible but INCORRECT ways to approach this problem and explain why they are wrong (Negative Constraints). 2. Then, solve it correctly avoiding these traps."*

4 Experimental Setups

4.1 Models and Datasets

To align with the fundamental study in the MAD literature (Du et al., 2023), we use GPT-3.5-Turbo (Schulman et al., 2022) as the debating and critic agent unless otherwise specified.

We use the GSM8K dataset (Cobbe et al., 2021), which contains graduate school mathematical reasoning tasks. To study the majority-error scenario, we extract 225 questions from the test set where a majority vote among three independently sampled agents is incorrect, which we refer to as *GSM-MajorityError*. To save token budgets, we evaluate the first 100 questions of *GSM-MajorityError* in all our experiments, which follows the common practice in prior works.

4.2 Evaluation of Naive MAD

To study the behavior of naive MAD on GSM-MajorityError, we vary the number of agents $A \in \{3,4,5\}$ and debate rounds $R \in \{1,2,3,4\}$ to examine how performance scales. Note that $R = 1$ corresponds to majority voting among A agents without debate.

We use the final *accuracy*, i.e., correctness after majority voting, as the primary evaluation metric, which extends to the two improved frameworks introduced in Section 3 and the prospective MAD with supervision framework to be discussed later.

Moreover, we examine the *flip statistics* of naive MAD. Let C and I denote being *correct* and *incorrect*, respectively. Using responses at the first ($R = 1$) and last ($R = 4$) rounds, we compute the number of agents that exhibit patterns of $C \to C$, $I \to C$, $C \to I$, and $I \to I$. These dynamics reveal whether interaction improves collective reasoning or amplifies collective error.

5 Results

In this section, we present empirical findings on how MAD behaves under majority-error conditions and evaluate whether our proposed extensions improve system stability and accuracy. We organize the results around four central questions introduced earlier.

5.1 Naive MAD Under Majority-Error Conditions

Across GSM-MajorityError, naive MAD shows limited ability to recover from an incorrect initial majority. We can observe from Figure 2 that as the number of debate rounds increases, accuracy fluctuates substantially rather than converging. In many cases, additional rounds even reduce accuracy, indicating that debate tends to amplify erroneous arguments instead of correcting them.

Accuracy does increase mildly as the number of agents grows, but this result cannot directly lead to the conclusion that MAD works, as sampling also contributes. As shown in Figure 2, most of the gains occur before any debate takes place.

In summary, the fluctuation of performance as debate rounds increase demonstrates the instability of naive MAD on difficult GSM-MajorityError problems. The improvement from invoking more agents without any debate at $R = 1$ further poses the question of whether the benefit comes from increased sampling or meaningful interaction among agents, which leads to the next section.

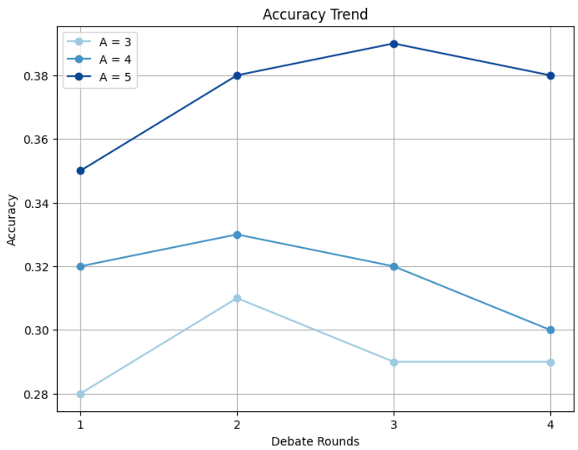


Figure 2: Accuracy trend of naive MAD across debate rounds with different numbers of agents.

5.2 Flip Dynamics During Debate

To understand why accuracy fluctuates across rounds and the source of improvement when increasing the number of agents, we examine how agents change their opinions across rounds, i.e., the flip dynamics as introduced in Section 4.

As shown in Figure 3, the correct-to-incorrect flips occur almost as frequently as the incorrect-to-correct flips. This symmetry indicates that debate does not reliably guide agents toward the correct answer–correct agents are just as easily persuaded in the wrong direction as incorrect agents are persuaded in the right one.

On the other hand, the number of agents that are initially correct and remain correct (correct-to-correct) increases steadily as the number of agents grows, reinforcing the observation that improvements come from sampling effects rather than the debate mechanism itself.

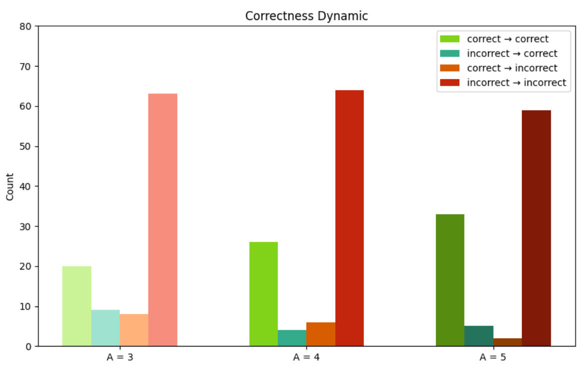


Figure 3: Opinion transition dynamics between the first and last round with different numbers of agents.

5.3 Is Repeated Sampling Better?

In prior works, comparisons between MAD and majority voting do not align on the *effective sample size*, i.e., the total number of times an LLM is sampled, which serves as an important metric of computational overhead. We define the effective sample size for MAD to be $A \times R$ and compare its performance with a repeated sampling baseline that conducts a majority vote among $N \times R$ independently sampled agents. As shown in Figure 4, repeated sampling outperforms naive MAD in almost all settings, further verifying that improvements attributed to debate largely result from increased sampling rather than the debate dynamics.

Although our evaluation is conducted on GSM-MajorityError, this finding potentially extends to broader cases. In fact, a recent study provides a general theoretical explanation. Choi et al. (2025) proposes a Dirichlet-Categorical Bayesian probabilistic model of the MAD framework and proves that each agent's belief in the correct answer actually follows a martingale process (Theorem 2): the expected belief of each agent across debate rounds remains unchanged. This means that MAD cannot effectively increase agents' belief in the correct answer. In contrast, the probability of getting the correct answer through majority vote increases exponentially with the number of agents (Theorem 1). Therefore, when the computation budget is controlled, majority voting over independent samples actually yields higher accuracy and avoids the fluctuations observed in debate.

Based on our observation and the theoretical insights, we argue that MAD's potential must come from its ability to identify and converge correct results through debates.

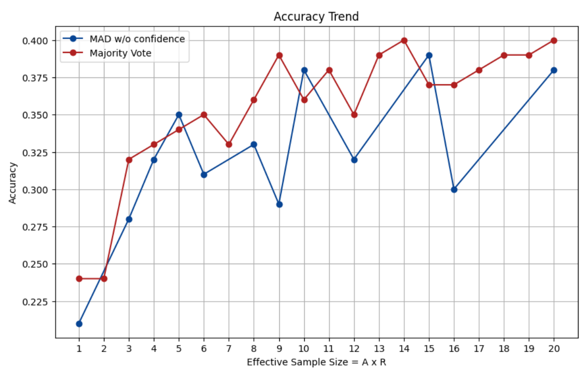


Figure 4: Comparison between naive MAD and repeated sampling under equal effective sample size ($A \times R$).

5.4 Improvements from Confidence Scoring and Role Specialization

We evaluate our proposed extensions to address the identified weaknesses of naive MAD. To some extent, these extensions follow the direction of identifying and converging to the correct answer as discussed in the previous section.

Confidence-Weighted Debate: By introducing an external critic to weight responses, the debate process is significantly stabilized. As shown in Figure 5, this mechanism effectively filters out minor fluctuations and reduces harmful persuasion from the incorrect majority, leading to consistent accuracy gains across debate rounds compared to the naive baseline.

Role Specialization: Utilizing the specialized roles defined in Section 3 enforces cognitive diversity and yields the strongest performance. This approach enables agents to approach the problem from complementary angles—reducing correlated errors and highlighting inconsistencies that homogeneous agents often overlook—thereby outperforming both naive MAD and the simple confidence-weighted approach.

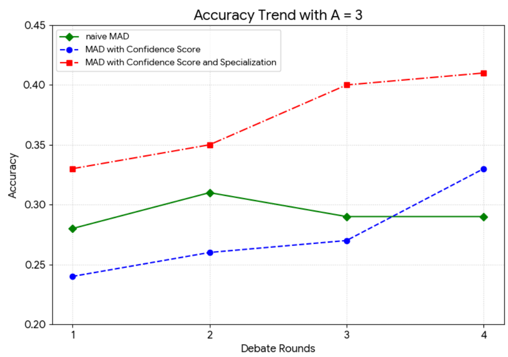


Figure 5: Accuracy trend under majority-error conditions with three debating agents (A = 3).

5.5 Naive MAD with High-Quality Supervision

Based on the previous observation, repeated sampling followed by majority voting often outperforms naive MAD, and confidence scoring and role specialization with agents of the same level bring limited help. This result drives us to explore another approach to unlocking the potential of MAD by seeking high-quality supervision that can effectively increase agents' belief in the correct answer.

Specifically, we invoke GPT-3.5-Turbo and GPT-5.1, respectively, to act as the supervisor/critic to simulate weak and strong supervision. The critics only help correct the logical errors or calculation mistakes and provide partial hints, assisting agents to converge to the correct answer, but not providing the answer directly.

As shown in Figure 6, empowered by a high-level critic (GPT-5.1), the performance of grows steadily and is improved to a large extent. By providing valuable feedback during the discussion, agents can realize their mistakes and correct themselves accordingly. Our findings indicate that MAD benefits significantly from high-quality supervision, such as human-AI cooperation.

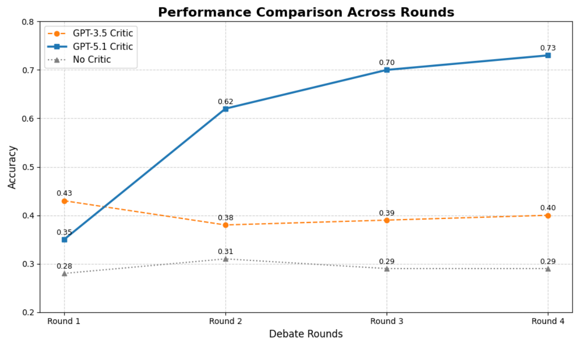


Figure 6: Performance comparison of naive MAD, supervision of the same level, and high-level supervision.

6 Limitations and Future Work

While this study offers critical insights into the instability of MAD under majority-error conditions, we acknowledge several limitations that define the scope of our findings and outline directions for future research.

Task Generalizability: Our evaluation focuses on arithmetic reasoning (GSM8K), where solutions are objectively verifiable. It remains an open question whether the observed dynamics—specifically the high rate of correct-to-incorrect flips—persist in open-ended or subjective tasks (e.g., creative writing), where consensus may be driven by rhetorical style rather than factual correctness.

Granularity of Interaction Analysis: Our current analysis relies on aggregate accuracy and flip statistics. We do not explicitly model the linguistic mechanisms of persuasion—such as why specific incorrect arguments successfully sway correct agents. A fine-grained analysis of argument quality and reasoning structure may offer further insights.

7 Conclusion

In this project, we study the majority error problem in MAD, where the majority of agents initially produce incorrect answers. By evaluating MAD's performance on a selected subset of mathematical reasoning tasks that simulate this setting, we reveal insights into the working mechanisms of MAD and explore approaches to improve its performance. Specifically, we find that naive MAD suffers from fluctuation across debate rounds and cannot effectively handle majority errors. Under the same computation budget, repeated sampling with majority voting actually outperforms naive MAD, which is supported by theoretical explanations. We argue that the potential of MAD must stem from effectively identifying the correct answer. For example, we demonstrate that confidence scoring and fine-grained role specialization improve the performance, and high-quality supervision potentially yields the best boost.

8 Team Contribution

All team members contributed substantially to the project. Below we summarize the primary responsibilities of each member.

Meitong Liu worked on running debate pipelines, analyzing majority-error behaviors, and discussing theoretical explanations.

Jieyi Zhao worked on analyzing majority-error behaviors and evaluating MAD's performance under external supervision.

Wangjia Zhan worked on experiments of the confidence score method and assisted in evaluating both improved debate frameworks.

Maojie Xu worked on running debate experiments, analyzing behaviors, and organizing most of the final report.

Ian Jiang worked on experiments of the role specialization method and assisted in evaluating both improved debate frameworks.

All members participated in weekly discussions, experiment planning, and the preparation of the final presentation and report.

9 Code Availability

All code used in this work is publicly available at: `https://github.com/nilgeoutim/CS546_MajorityErrorDebate`

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *Preprint*, arXiv:2309.13007.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Zijie Lin and Bryan Hooi. 2025. Enhancing multi-agent debate system performance via confidence expression. *Preprint*, arXiv:2509.14034.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *Preprint*, arXiv:2310.10076.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, and 1 others. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2(4).

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. *Preprint*, arXiv:2406.07791.

Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 20090–20111. Association for Computational Linguistics.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *Preprint*, arXiv:2509.05396.