FREE-MAD: Consensus-Free Multi-Agent Debate

Yu Cui[1] **Hang Fu**[1] **Haibin Zhang**[2,3]* **Licheng Wang**[1] **Cong Zuo**[1]*
[1]Beijing Institute of Technology
[2]Yangtze Delta Region Institute of Tsinghua University, Zhejiang
[3]Jiaxing Key Laboratory of Artificial Intelligence and Cyber Resilience
cuiyu@bit.edu.cn, bchainzhang@aliyun.com

Abstract

Multi-agent debate (MAD) is an emerging approach to improving the reasoning capabilities of large language models (LLMs). Existing MAD methods rely on multiple rounds of interaction among agents to reach consensus, and the final output is decided by majority voting in the last round. However, this consensus-based design faces several limitations. First, multiple rounds of communication increases token overhead and limits scalability. Second, due to the inherent conformity of LLMs, agents that initially produce correct responses may be influenced by incorrect ones during the debate process, causing error propagation. Third, majority voting introduces randomness and unfairness in the decision-making phase, and can degrade the reasoning performance. To address these issues, we propose FREE-MAD, an alternative and novel MAD framework that eliminates the need for consensus among agents. FREE-MAD introduces a novel score-based decision mechanism that evaluates the entire debate trajectory rather than relying on the last round only. This mechanism tracks how each agent's reasoning evolves, enabling more accurate and fair outcomes. In addition, FREE-MAD reconstructs the debate phase by introducing anti-conformity, a mechanism that enables agents to mitigate excessive influence from the majority. Experiments on eight benchmark datasets demonstrate that FREE-MAD significantly improves reasoning performance while requiring only a single-round debate and thus reducing token costs. We also show that compared to existing MAD approaches, FREE-MAD exhibits improved robustness in real-world attack scenarios.

1 Introduction

Large language models (LLMs), due to their strong reasoning capabilities, have been widely applied in domains such as chatbots (Li et al., 2024a), programming (Yan et al., 2024), healthcare (Liu et al., 2024a), and cybersecurity (Zou et al., 2024). Recent applications of LLM agents (Luo et al., 2025b) have placed increasing demands on their reasoning performance. To improve the reasoning accuracy of LLM agents, many studies have explored training-free methods such as Chain-of-Thought (CoT) (Wei et al., 2022), self-refinement (Madaan et al., 2023) and self-consistency (Wang et al., 2023). However, these methods focus on the usage of single LLMs and lack collaboration among multiple models. In contrast, multi-agent debate (MAD) (Zeng et al., 2025; Du et al., 2024; Chan et al., 2024; Liu et al., 2025) has emerged as a prominent solution, showing that multiple agents engaged in multi-round interactions can achieve substantially better reasoning performance than a single agent. Indeed, MAD has been used in various scenarios such as software issue resolution (Li et al., 2025), mathematical reasoning (Zhang and Xiong, 2025), and code summarization (Chun et al., 2025).

Existing efforts to optimize MAD focus primarily on the reasoning strategies of individual agents during debates (Liu et al., 2025) or improving scalability (Zeng et al., 2025). Meanwhile, recent studies reveal that LLM agents can exhibit conformity (Weng et al., 2025; Zhu et al., 2025; Cho et al., 2025), meaning that during multi-agent interactions, agents tend to favor answers endorsed by the majority. In existing MAD frameworks, conformity-driven consensus (Sun et al., 2024; Zeng et al., 2025; Li et al., 2024b) is employed during the debate process to obtain the correct answer as the final decision (Chan et al., 2024). However, such consensus reduces reasoning accuracy. The consensus-based MAD schemes suffer from the Silent Agreement problem (Wang et al., 2025b). Even when the agents start with divergent opinions, they remain silent during the discussion due to conformity. As a result, the agent group eventually

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 31977–31997

provides an incorrect answer. More importantly, consensus naturally demands more debate rounds, increasing token consumption and limiting scalability.

To address these limitations, we propose FREE-MAD, a consensus-free MAD framework that reconstructs both the debate stage and the decision stage. In the debate stage, we integrate the conventional conformity mode as used in prior work with a new mode called anti-conformity. In particular, the anti-conformity mode leverages CoT to encourage agents to identify flaws in the outputs from other agents; this situation is in contrast to existing approaches that use consensus as an indicator of correctness. In the decision stage, we propose a score-based mechanism that evaluates *all* intermediate outputs across debate rounds, instead of focusing solely on the final round as in traditional MAD frameworks. By tracking changes in the reasoning trajectories of all agents, FREE-MAD assigns scores to all candidate responses without requiring consensus in the debate stage. Furthermore, we theoretically formalize the MAD protocol, enabling a formal comparison between FREE-MAD and existing MAD variants.

To evaluate FREE-MAD, we have conducted extensive experiments on eight benchmark datasets, covering knowledge-based reasoning, logical reasoning, and mathematical reasoning with varying levels of difficulty. We show that FREE-MAD outpaces baseline approaches in terms of reasoning accuracy. In particular, FREE-MAD achieves improved accuracy with fewer debate rounds, thus accelerating the debate process while maintaining strong scalability. Furthermore, we show that FREE-MAD exhibits enhanced robustness (Chen et al., 2024a) and maintains accurate reasoning against attacks (He et al., 2025), where interactions among agents might be partially disrupted. We summarize our contributions as follows:

- We propose a novel consensus-free MAD framework, called FREE-MAD, with dedicated optimizations in both the debate and decision stages. We design a score-based decision mechanism that evaluates all intermediate results across debate rounds, enabling accurate reasoning without requiring consensus.
- We provide a syntax for the MAD protocol and perform a formal analysis of FREE-MAD and existing mechanisms.
- We implement FREE-MAD and conduct extensive experiments on eight benchmarks. We show that FREE-MAD outperforms existing protocols in terms of reasoning accuracy, scalability, and robustness.

2 Related Work

MAD Protocol Security. In the debate stage, traditional consensus-oriented approaches for agents tend to lack robustness in unreliable network environments. Communication attacks (He et al., 2025) can force some agents to withdraw from the debate, preventing them from receiving others' responses. This delays consensus and increases overhead in adaptive schemes (Liang et al., 2024), ultimately reducing reasoning accuracy. Another line of work allows agents to return both their individual responses and a self-assessed confidence score (Chen et al., 2024b), which is then used in a weighted aggregation of the final result.

However, due to LLM hallucinations (Ji et al., 2023), such confidence may be unreliable. Existing decision mechanisms in MAD are inadequate for addressing the security risks associated with LLM conformity. In real-world deployments, if a small subset of agents is compromised via prompt injection attacks (Greshake et al., 2023; Liu et al., 2024b; Zhan et al., 2025), the system may converge toward a shared but incorrect answer. When decisions are made using mechanisms such as majority voting, this can lead to a complete failure of the MAD system. Other approaches (Liang et al., 2024) use an LLM-as-a-Judge (Zheng et al., 2023) framework, where an LLM decides the final outcome. This approach can produce biased results and is prone to conformity, making it effectively equivalent to majority voting. In addition, if the LLM is compromised by a prompt injection attack, the accuracy of the MAD system's final output may drop significantly.

Agent Diversity in MAD. In previous studies, although some MAD frameworks have considered heterogeneity and diversity among agents, their experimental evaluations were typically conducted using LLMs with similar model sizes (Yang et al., 2025), such as Llama3.1-8B[1], Qwen2.5-7B[2], and Gemma-2-9B[3]. This setup significantly limits the effectiveness of MAD and, in certain cases, re-

[1]https://huggingface.co/meta-llama/Llama-3.1-8B
[2]https://huggingface.co/Qwen/Qwen2.5-7B
[3]https://huggingface.co/google/gemma-2-9b

sults in worse performance than self-consistency approaches.

In real-world deployments, however, we aim to enable collaboration among diverse LLMs, each possessing different strengths (Chan et al., 2024; Liu et al., 2025), to act as equal peers within a MAD framework. Such collaboration is expected to outperform the strongest single agent on the same task and to solve a broader range of problems through cooperation, analogous to human team-based collaboration.

However, when the participating agents in a heterogeneous MAD system exhibit substantial differences in their capabilities, it may lead to significant variance in their confidence regarding their own responses. As a consequence of conformity in LLMs (Weng et al., 2025), agents may tend to adopt the outputs of peers, even when those outputs are incorrect. While conformity can be beneficial in facilitating consensus, it also introduces detrimental effects that require external mitigation.

MAD Protocols. When MAD was introduced, it was intended to improve reasoning by encouraging LLMs to reach consensus across their answers. However, such a consensus cannot be guaranteed in theory and is generally achievable only in practice (Du et al., 2024). Moreover, consensus becomes easier to achieve when agents adjust their trust between self-generated and externally generated content, yet this adjustment tends to reduce reasoning accuracy. The researchers did not investigate the deeper conformity issue that underlies this performance drop. Subsequent studies have focused on optimizing the debate stage of MAD (Chan et al., 2024; Chen et al., 2024b; Zeng et al., 2025; Liang et al., 2024; Liu et al., 2025). In these works, consensus is commonly treated as the default goal of the debate stage (Li et al., 2024b). In this paper, we show that MAD can operate effectively without requiring consensus.

Consensus in MAD. Reaching consensus was the fundamental objective when MAD was first proposed (Du et al., 2024). It is also a necessary condition for obtaining a correct final answer. To the best of our knowledge, all existing MAD methods adopt consensus as a core mechanism in their underlying design (Li et al., 2024b). Chan et al. (2024) does not explicitly require agents to reach consensus during the debate stage. However, it still applies majority voting in the decision phase, which preserves the essential logic of consensus.

3 Preliminary Analysis

3.1 MAD Protocol: A Formal Two-Phase Decomposition

To enable a formal analysis, we decompose the MAD protocol into two core stages: **Debate** and **Decision**. The Debate stage internally unfolds over R iterative rounds, culminating in a set of final answers. Formally, given a set of N agents denoted as $\{a_i\}_{i=1}^{N}$, the protocol is defined as:

$$\{r_i^R\}_{i=1}^{N} \leftarrow \texttt{Debate}(\{a_i\}_{i=1}^{N},\ q,\ p,\ R), \quad (1)$$

$$r_{\text{final}} \leftarrow \texttt{Decide}(\{r_i^R\}_{i=1}^{N}). \quad (2)$$

In the **Debate** stage, all agents engage in a multi-round interaction based on a user query q and a guiding prompt p that specifies how agents should debate. The debate unfolds over R rounds. The debate begins with an initial step where each agent generates a preliminary response to q, which is then broadcast to all other agents as an auxiliary context (Yang et al., 2025). Subsequently, each agent a_i iteratively updates its own answer r_i^k over R rounds, resulting in a final set of responses $\{r_i^R\}_{i=1}^{N}$.

The complete history of utterances up to round $k-1$ is denoted as the context $C^{(k-1)}$. We model LLM agents as a probabilistic process to capture their generative behavior. In round k, agent a_i produces its response r_i^k by sampling from a conditional probability distribution defined over the preceding context and p:

$$r_i^k \sim P_{a_i}(r|C^{(k-1)}, p). \quad (3)$$

To study the interplay between independent reasoning and conformity, we model the overall probability distribution P_{a_i} with a formulation that separates the contributions of the two factors:

$$\begin{aligned} P_{a_i}(r|C^{(k-1)}, p) = \frac{1}{Z} \cdot P_{\text{in}}(r|q,p) \\ \cdot \exp(\beta(p) \cdot S_{\text{con}}(r, C^{(k-1)})), \end{aligned} \quad (4)$$

where the independent reasoning distribution $P_{\text{in}}(r \mid q,p)$ mathematically characterizes the agent's intrinsic reasoning ability given the question q and prompt p. This ability excludes peer influence. The conformity score $S_{\text{con}}(r, C^{(k-1)})$ measures how much a candidate response r aligns with peer utterances in $C^{(k-1)}$. Its effect is scaled by the conformity parameter $\beta(p)$, which is determined by p. By default, LLMs exhibit a tendency to

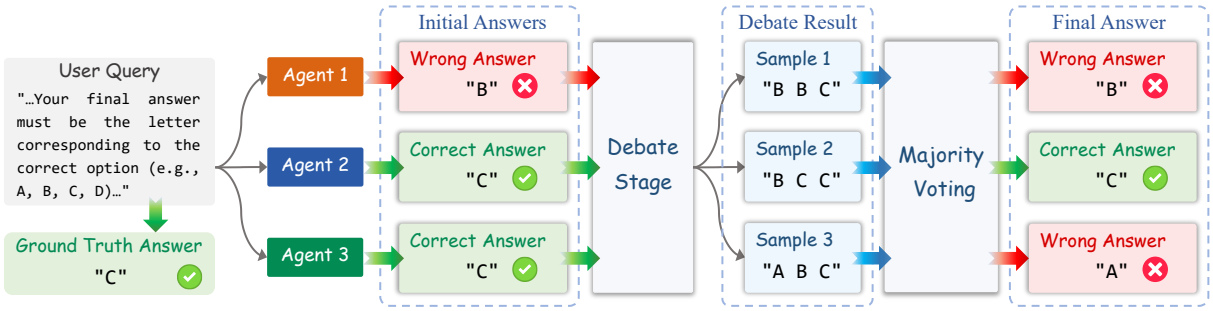


Figure 1: Existing MAD approaches may obtain final answers that are even less accurate than the initial ones.

conform. We capture this by assuming $\beta_{\text{default}} > 0$ when p is empty or neutral. If p encourages critical reasoning, $\beta(p)$ can be negative, acting as a regularizer that reduces alignment and pushes responses toward P_{in}. Conversely, if p encourages agreement with majority opinions, $\beta(p)$ assumes a larger positive value, which speeds up convergence towards consensus. Finally, the model combines all components via an exponential transformation and normalizes by Z to ensure a valid probability distribution.

In the **Decision** stage, a final output r_{final} is selected from $\{r_i^R\}_{i=1}^N$, typically through mechanisms such as majority voting. In this work, R excludes initial response generation and begins once agents start receiving responses from others.

3.2 Weaknesses of Existing MAD Approaches

Reasoning Accuracy. MAD approaches (Du et al., 2024; Yang et al., 2025; Li et al., 2024b) design the decision stage to operate on the final round's N responses in the debate, while overlooking the remaining $R \times N$ intermediate responses that emerged throughout the debate process. This omission diminishes the influence of these earlier responses on determining r_{final}, thereby reducing both the accuracy and fairness of the final outcome. In consensus-based debates, the process ends once the agents reach agreement, even if the answer is incorrect (Chen et al., 2024b; Wang et al., 2025b).

However, Du et al. (2024) has shown that it is still possible for the correct answer to emerge during later stages of the debate, even if none of the agents initially generates a correct answer. Early termination thus reduces MAD's problem-solving accuracy. From empirical observations, we find that the initial responses generated independently by multiple agents may outperform the debate results obtained after applying MAD. As shown in Figure 1, applying majority voting directly to initial answers can yield the correct result, while debate outcomes may be incorrect. We illustrate three possibilities. In Sample 1, the agents reach consensus on an incorrect answer. Sample 2 represents the desired outcome, where the correct answer "C" holds the majority. Outcomes such as Sample 3 have received little attention. The set $\{r_i\}_{i=1}^N$ contains entirely distinct outputs, with no repetitions or equal counts for multiple answers. Under these circumstances, the final answer is determined either by selecting a response at random from the set $\{r_i\}_{i=1}^N$ or by choosing the first one. Both strategies substantially degrade the accuracy of MAD. Therefore, majority voting is unsuitable for decisions based on debate outcomes. More robust and practical mechanisms are required.

Robustness. Most existing multi-agent systems lack robustness (Chen et al., 2024a; Zhang et al., 2024; He et al., 2025). MAD is even more vulnerable to attacks (Qi et al., 2025) due to its consensus mechanism. Researchers (Luo et al., 2025a) mainly enhance the robustness of multi-agent systems by introducing blockchain and leveraging its traditional distributed consensus protocols. However, the use of blockchain, including smart contracts (Li et al., 2023) and consensus execution (Zhang et al., 2023; Duan et al., 2018), drastically impacts system performance.

4 Methodology

In this section, we first formalize FREE-MAD. We then present two core techniques: a consensus-free debate protocol and a score-based decision mechanism. Finally, we analyze FREE-MAD in comparison with existing approaches.

4.1 FREE-MAD

FREE-MAD focuses on the complete set of outcomes generated throughout the entire MAD process, rather than limiting attention to only the final-

round responses, as done in traditional approaches. Our proposed debate protocol incorporates all responses into the decision-making process. This perspective can be formally represented using the following matrix formulation:

$$\texttt{Decide}[\alpha_0, \alpha_1, \ldots, \alpha_{R-1}, \alpha_R] = \begin{bmatrix} r_1^0 & r_1^1 & \cdots & r_1^{R-1} & r_1^R \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_N^0 & r_N^1 & \cdots & r_N^{R-1} & r_N^R \end{bmatrix} \rightarrow r_{\text{final}}. \quad (5)$$

The highlighted entries ($\alpha_0, \alpha_1, \ldots, \alpha_{R-1}$) represent components that were not considered in previous MAD approaches. r_i^0 denotes the initial response generated by agent i. Unlike prior MAD frameworks that decouple debate and decision stages, FREE-MAD innovatively integrates decision-making into the debate phase, resulting in a unified process (see Figure 2).

4.2 Consensus-Free Debate

The agents in MAD exchange responses and update their answers based on a prefix prompt. A common prompt is "The responses from other agents are as follows" (Du et al., 2024). Due to the conformity of LLMs, such prompts often bias the agent toward the majority answer. This undermines rigorous reasoning and raises the likelihood that reasoning errors remain unnoticed. To mitigate this conformity, we design an additional CoT-based prompt (see Appendix E) appended after the prefix prompt, motivated by (Weng et al., 2025). We provide agents with a scenario that includes adversarial agents. We instruct agents to carefully assess the discrepancies between their own answers and those from peers. Agents are expected to change their beliefs only if there is a clear indication that their own answer is incorrect, rather than aiming to reach consensus with others. This mechanism is intended to reduce the propagation of incorrect answers during the MAD process.

Our structured and critical reasoning prompt forms the core mechanism for optimizing the probabilistic model (Equation 4) and has two main effects. First, it improves the quality of the independent reasoning distribution P_{in}. FREE-MAD requires each agent to provide an answer along with a detailed reasoning trace, which is incorporated into the next-round context $C^{(k-1)}$. Agent a_i uses this context to analyze peers' reasoning rather than just their answers. If the reasoning behind the majority answer is flawed, a_i's own critical thinking will assign a very low probability to that answer in P_{in}. Second, it balances conformity. Even if a popular but incorrect answer has a high S_{con}, a low P_{in} keeps its overall probability P_{a_i} low.

Algorithm 1: MAD Protocol via Score-Based Decision

Input: Answer matrix $A \in \mathbb{R}^{N\times(R+1)}$ from N agents over R rounds; Task input q; Weights $\mathcal{W} = \{w_i\}_{i=1}^4$; Guiding prompt p

Output: Final answer

Score dictionary $S \leftarrow \emptyset$
for $k \leftarrow 0$ **to** R **do**
  $f = (k+1)^{-1}$
  for $i \leftarrow 1$ **to** N **do**
    if $k \neq 0$ **then**
      Context $C \leftarrow$ responses in $k-1$
    $r_i^k \leftarrow P_{a_i}(q, p)$; Update C and A with r_i^k
    $\hat{r} \leftarrow A[i][k]$
    if $k = 0$ **then**
      $S[\hat{r}] \leftarrow S[\hat{r}] + w_1 f$
    else
      $r_p \leftarrow A[i][k-1]$
      if $\hat{r} \neq r_p$ **then**
        if $r_p \in S$ **then**
          $S[r_p] \leftarrow S[r_p] - w_2 f$
        $S[\hat{r}] \leftarrow S[\hat{r}] + w_3 f$
      else
        $S[\hat{r}] \leftarrow S[\hat{r}] + w_4 f$
Remove invalid keys from S.
$MS \leftarrow \arg\max_k S[k]$
if $|MS| > 1$ **then**
  Randomly choose $r_{\text{final}} \in MS$
else
  $r_{\text{final}} \leftarrow MS[0]$
Return r_{final}

4.3 Score-Based Decision Mechanism

Based on the existing MAD framework (Du et al., 2024) and formalization of the traditional MAD protocol (Subramaniam et al., 2025), we describe our protocol as shown in Algorithm 1. This method maintains a matrix $A \in \mathbb{R}^{N\times(R+1)}$ that records the real-time responses of each agent across debate rounds. Concurrently, a score dictionary S is used

Approaches	Decision	Anti-Conformity	Fairness	Security
Sparse MADLi et al. (2024b)	Majority Voting	✗	✗	✗
Liang et al. (2024)	LLM-as-a-Judge	✗	✗	✗
ReConcile (Chen et al., 2024b)	Weighted Voting	✗	✓	✗
ChatEval (Chan et al., 2024)	Majority Voting	✗	✗	✗
DMAD (Liu et al., 2025)	Majority Voting	✗	✗	✓
SoM (Du et al., 2024)	Majority Voting	✗	✓	✓
FREE-MAD-N	score	✓	✓	✓
FREE-MAD-C	score	✗	✓	✓

Table 1: A comprehensive comparison between our proposed approach and existing methods.

to track the scores associated with multiple answers that emerge throughout the debate. The mechanism evaluates the likelihood of an answer being correct by tracking whether agents exhibit a shift in their opinions across rounds (whether the answer provided in the current round differs from that in the previous round). The agents in this framework are not designed to seek consensus; instead, they rigorously assess the reasoning behind the answers. Therefore, a change in an agent's response is interpreted as an indicator that a more accurate answer has been identified. Specifically, answers that agents abandon are considered more likely to be incorrect, whereas newly adopted answers are treated as more likely to be correct. This dynamic informs the score updates within the dictionary S.

We assign different weights $w_i f$ to answers based on whether agents have changed their responses between rounds. Here, $w_i \in \mathcal{W}$ represents system-defined parameters, and f is a correction factor inversely proportional to the current round number. As the number of rounds increases, the amount of contextual information each agent receives also grows, thereby increasing the risk of conformity, particularly for agents based on smaller models. To limit the influence of conformity, the impact of opinion shifts in later rounds is down-weighted via the factor f. We define MS as the set of answers with the highest score. Although this set usually contains a single candidate, we adopt a randomized selection strategy to maintain theoretical robustness.

4.4 Framework Design

As noted in (Weng et al., 2025), LLMs' conformity presents a double-edged sword. On the one hand, it fosters consensus and cohesive outcomes. On the other hand, it undermines the reliability of agents' judgments in sensitive domains such as voting. Debate based on anti-conformity mitigates the negative effects of blind conformity among agents. However, for relatively simple tasks, LLMs may show excessive anti-conformity, which causes stubbornness and reduces reasoning accuracy. Therefore, we argue that conformity-oriented and anti-conformity-oriented debate modes should be adapted and switched according to the task. To achieve finer control over excessive conformity, the weight parameter $\mathcal{W}$ in the score mechanism can be adjusted, which regulates the balance between an agent changing its opinion and maintaining its original stance. Based on this idea, we propose FREE-MAD, which consists of two sub-schemes that share a score-based decision stage. FREE-MAD-N integrates an anti-conformity debate process, while FREE-MAD-C integrates a conformity-based debate process. Together, the schemes extend the framework's applicability to a wider range of real-world scenarios.

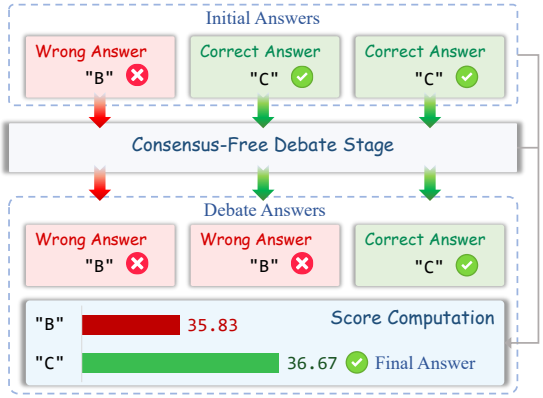


Figure 2: The inference process of our MAD framework. When the correct answers are in the minority in the final round, the framework is still able to identify the correct one as the final answer.

4.5 Analysis

We compare our proposed FREE-MAD framework with existing MAD approaches, as shown in Table 1. Below, we provide a detailed analysis of the advantages of FREE-MAD.

Accuracy. The FREE-MAD framework adopts a consensus-free debate strategy, which helps reduce the influence of conformity. This approach encourages agents to concentrate more on the correctness of reasoning steps, and may alleviate the effect of error propagation. In addition, existing methods usually require multiple rounds of debate to reach consensus. When the number of rounds R is small, such as when $R = 1$, these approaches may experience reduced reasoning performance due to the lack of consensus, which can lead to lower accuracy. In contrast, the performance of FREE-MAD is not closely tied to R.

Scalability. Prior work (Zeng et al., 2025) provides

a general complexity analysis of the token cost (TC) in MAD as: $\mathcal{O}\big(NR^2V + N^2RV + NR|q|\big)$, where V is the maximum token cost for each agent. Therefore, a practical MAD framework should aim to achieve high accuracy with fewer agents and fewer rounds (Liu et al., 2025). The consensus-building process generally takes 2 to 3 rounds (Chen et al., 2024b; Du et al., 2024; Yang et al., 2025; Xiong et al., 2023) to be effective. In contrast, our method theoretically requires only a single round of debate without any consensus constraint, which substantially reduces token consumption.

Security. Under communication attacks, agents that withdraw from the debate process generally trigger only $S[\hat{r}] \leftarrow S[\hat{r}] + w_4 f$ in Algorithm 1 under our score scheme, because these agents retain context containing only their own prior responses, which does not affect the overall debate process. In addition, the score strategy is executed entirely outside the LLM reasoning and follows a deterministic protocol, rendering it immune to LLM hallucination.

Fairness. During the debate stage, some approaches adopt role-based debate strategies (Chan et al., 2024), where agents are assigned unequal statuses and perform different functions. This design reinforces the implicit biases of LLMs (Vasista et al., 2025; Myung et al., 2025; Kim et al., 2024) and undermines the fairness of MAD systems (Xiong et al., 2023). In contrast, agents in FREE-MAD do not require any predefined roles, and they participate equally in the debate process.

5 Experiments

5.1 Experimental Setup

Evaluation Benchmark. Based on the comparison in Table 1 and the analysis in Section 4.5, we select the SoM framework (Du et al., 2024) as the baseline for our experiments to ensure a fair comparison (SoM is also widely adopted as a baseline in related work (Chen et al., 2024b; Wang et al., 2025a; Li et al., 2024b)). Moreover, our proposed FREE-MAD is implemented on top of SoM to minimize the influence of confounding factors. In Algorithm 1, the weights $\mathcal{W}$ are initialized to $\{20, 25, 30, 20\}$ based on theoretical analysis. We present the experimental setup for evaluating the security of FREE-MAD in Appendix D. For the ablation study, we compare four schemes, as summarized in Table 2, including our FREE-MAD and the SoM baseline. This comparison highlights the effectiveness of the two core modules we developed.

Schemes	FREE-MAD-N	FREE-MAD-C	Baseline 1	Baseline 2 (SoM)
Debate Decision	Anti-conformity Score	Conformity Score	Anti-conformity Majority Voting	Conformity Majority Voting

Table 2: Module configurations of multiple comparative variants in ablation experiments.

Datasets. To comprehensively evaluate the capability of FREE-MAD, we conduct experiments on 8 benchmark datasets. For mathematical reasoning, we use GSM-Ranges (Shrestha et al., 2025) (levels 4 and 6), AIME2024, AIME2025 (Art of Problem Solving, 2025), and MATH500 (Lightman et al., 2024). For logical reasoning, we employ StrategyQA (Geva et al., 2021) and the Logical Fallacies dataset of MMLU (Hendrycks et al., 2021). For knowledge and theoretical reasoning, we adopt the multiple-choice questions dataset from AICrypto (Wang et al., 2025c), which constitutes the first benchmark specifically constructed to assess the cryptographic capabilities of LLMs.

Agent Groups. To ensure that the MAD framework possesses the basic capability to handle our datasets, we design two configurations of MAD. For AIME2024 and AIME2025, we construct MAD with $N = 3$ based on Qwen1.5-7B-Chat[4] and DeepSeek-V3 (DeepSeek-AI et al., 2024). For the other datasets, we uniformly construct MAD with $N = 4$ using Qwen1.5-7B-Chat and Qwen2.5-72B-Instruct (Qwen et al., 2025). Other details are provided in Appendix.

5.2 Evaluation Metrics

To evaluate the reasoning performance and scalability of MAD, we assess both inference accuracy and token consumption. Following (Zeng et al., 2025), we adopt token consumption as the metric for scalability. The computation of accuracy follows Algorithm 1, while the calculation of token consumption is defined as follows:

$$\text{TC} = \sum_{k=0}^{R} \sum_{i=1}^{N} \mathcal{T}_k^i, \tag{6}$$

where $\mathcal{T}_k^i$ denotes the number of output tokens generated by agent a_i in the k-th round.

6 Main Results and Discussion

6.1 Reasoning Performance

The evaluation results of reasoning accuracy on eight benchmarks are presented in Figure 3 ($R = 1$)

[4]https://qwenlm.github.io/zh/blog/qwen1.5

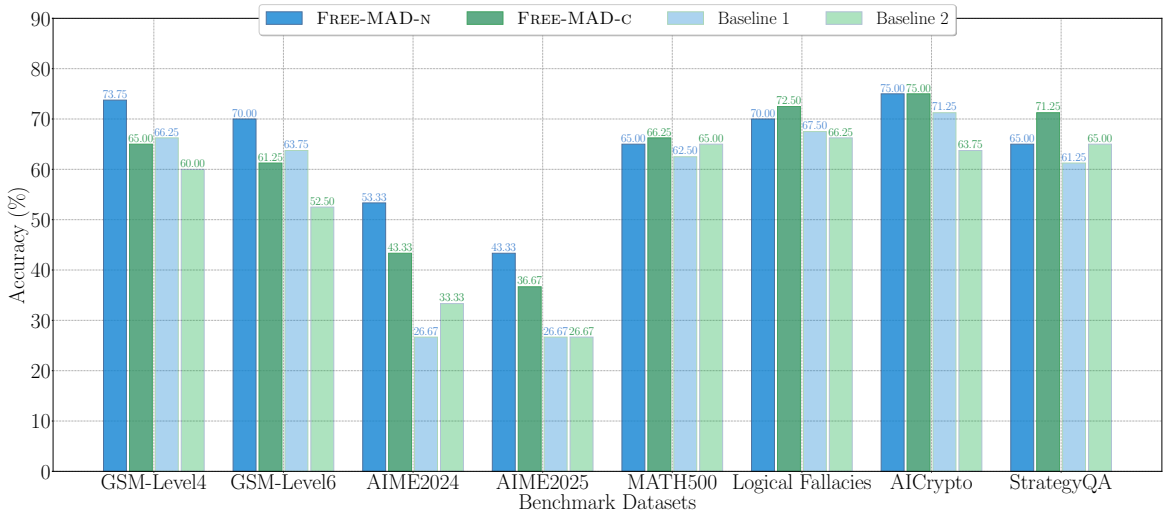
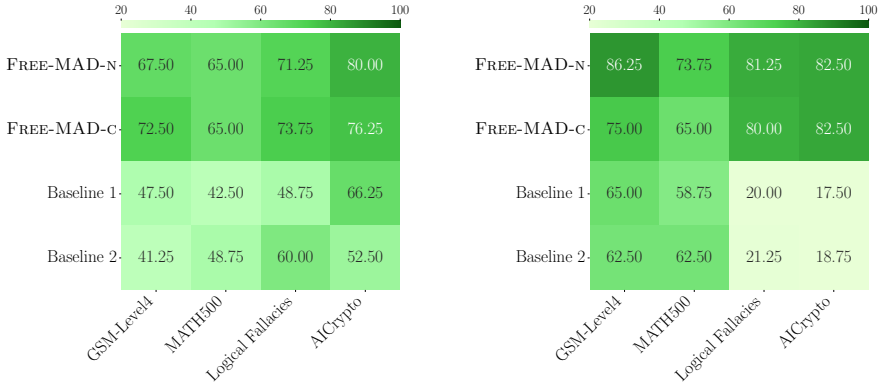


Figure 3: Comprehensive comparative experimental results for MAD frameworks across multiple benchmarks.



(a) Experimental results when $R = 1$.

(b) Experimental results when $R = 2$.

Figure 4: Empirical evaluation of the security of MAD frameworks across multiple benchmarks, showing the comparison of their reasoning accuracy under communication attacks across varying numbers of rounds.

and Table 4 ($R = 2$). Overall, our proposed FREE-MAD substantially outperforms the baselines, achieving average improvements of 13.0% and 16.5% over baselines, respectively. These results demonstrate a significant enhancement in reasoning performance. In particular, for mathematical reasoning tasks, the advantage of FREE-MAD becomes more evident with increasing problem difficulty. Reducing conformity clearly improves the effectiveness of MAD. The specific effects of conformity and anti-conformity on the reasoning process of LLMs are detailed in Appendix F. Notably, under anti-conformity, the reasoning process of LLMs appears to be more rational.

On the MATH500 dataset, we observe that weaker models exhibit a previously mentioned tendency toward rigidity in reasoning when conformity is suppressed. In such cases, these models fail to switch to correct lines of reasoning, resulting in comparable performance between FREE-MAD and the baselines. This limitation is expected, as a fixed set of agents cannot be universally optimal across all task categories. For logical and knowledge-based reasoning, FREE-MAD still significantly outperforms the baselines. However, FREE-MAD-C consistently achieves better results than FREE-MAD-N, because for models lacking relevant knowledge, idea switching under anti-conformity tends to occur with relatively high randomness. Consequently, in such scenarios, conformity may lead to more effective outcomes. By comparing the four variants in our ablation study, we demonstrate that the proposed core score-based decision mechanism exhibits superior performance.

Schemes	FREE-MAD-N	FREE-MAD-C	Baseline 1	Baseline 2
Accuracy	64.43% (↑ 16%/19%)	61.41% (↑ 10%/14%)	55.73%	54.06%

Table 3: Comparison of reasoning accuracy between our schemes and baselines when $R = 1$.

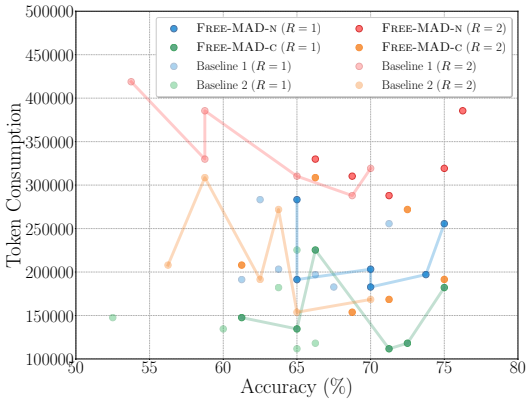


Figure 5: Comparison of token consumption and reasoning accuracy between our proposed schemes and baselines under different debate rounds.

6.2 Scalability

We compared the token consumption and reasoning accuracy of four MAD variants under $R = 1$ and $R = 2$, as shown in Figure 5. With an increasing number of debate rounds, the reasoning accuracy of MAD improves. Notably, FREE-MAD achieves accuracy comparable to or even higher than the two-round baseline 2 setting with only a single debate round, while incurring almost no additional token consumption, demonstrating promising scalability. Specifically, FREE-MAD-N achieves stronger reasoning accuracy compared to the baselines, whereas FREE-MAD-C offers better scalability. More importantly, our approach eliminates the need for multiple debate rounds to reach consensus, which significantly reduces the execution time of the MAD system.

6.3 Security

Compared to the normal scenario, both baseline 1 and baseline 2 exhibit a substantial drop in accuracy, reaching up to 20% (see Figure 4). In contrast, FREE-MAD consistently maintains very high accuracy. Interestingly, in some cases, it even slightly outperforms the original accuracy. This behavior can be attributed to the fact that communication attacks prevent some agents from receiving responses from others, while simultaneously reducing the probability of receiving incorrect information. These results demonstrate that FREE-MAD possesses strong robustness and security.

7 Conclusion

This paper proposes FREE-MAD, a novel MAD framework that integrates controllable conformity with a score-based decision mechanism. Unlike traditional MAD approaches, FREE-MAD does not rely on multi-round interactions or need to reach a consensus. By evaluating the entire debate trajectory instead of relying solely on the final round, the accuracy of deciding the answer is enhanced. Extensive experiments show that FREE-MAD outperforms existing MAD approaches in terms of reasoning accuracy, scalability, and robustness.

8 Limitations

In this paper, we construct a general MAD framework that incorporates a controllable conformity debate process together with a score-based mechanism that determines the final decision across all debate outcomes. Although FREE-MAD achieves significant reductions in overhead and improvements in accuracy, its token overhead still affects practical deployment compared with single model inference, and further optimization is necessary. FREE-MAD provides a general framework for multiple MAD schemes. Evaluating optimization gains across more schemes is necessary, but resource constraints prevent such evaluations.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62272043), Beijing Natural Science Foundation (Grant No. L251002), National Natural Science Foundation of China (Grant No. 62372040, 62572045), Natural Science Foundation of Beijing, China (Grant No. 4254086) and Yangtze Delta Region Institute of Tsinghua University, Zhejiang (No.LZZLX24F007).

We thank Jonathan Santilli for providing an industrial grade and robust Python implementation of FREE-MAD[5]. This implementation enhances the practical applicability of FREE-MAD and mitigates inherent limitations of the original method.

References

Art of Problem Solving. 2025. AIME Problems and Solutions. Accessed: 2025-05-15.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

[5]https://github.com/jonathansantilli/freemad

Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. 2024a. Blockagents: Towards byzantine-robust llm-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award Celebration Conference - China 2024*, ACM-TURC '24, page 187–192, New York, NY, USA. Association for Computing Machinery.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Young-Min Cho, Sharath Chandra Guntuku, and Lyle Ungar. 2025. Herd behavior: Investigating peer influence in llm-based multi-agent systems. *Preprint*, arXiv:2505.21588.

Jina Chun, Qihong Chen, Jiawei Li, and Iftekhar Ahmed. 2025. Is multi-agent debate (mad) the silver bullet? an empirical analysis of mad in code summarization and translation. *arXiv preprint arXiv:2503.12029*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Sisi Duan, Michael K. Reiter, and Haibin Zhang. 2018. Beat: Asynchronous bft made practical. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 2028–2041, New York, NY, USA. Association for Computing Machinery.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 79–90, New York, NY, USA. Association for Computing Machinery.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengfei He, Yuping Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. 2025. Red-teaming LLM multi-agent systems via communication attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6726–6747, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. DEBATE: Devil's advocate-based assessment and text evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1885–1897, Bangkok, Thailand. Association for Computational Linguistics.

Han Li, Yuling Shi, Shaoxin Lin, Xiaodong Gu, Heng Lian, Xin Wang, Yantao Jia, Tao Huang, and Qianxiang Wang. 2025. Swe-debate: Competitive multi-agent debate for software issue resolution. *arXiv preprint arXiv:2507.23348*.

Huizhong Li, Yujie Chen, Xiang Shi, Xingqiang Bai, Nan Mo, Wenlin Li, Rui Guo, Zhang Wang, and Yi Sun. 2023. Fisco-bcos: An enterprise-grade permissioned blockchain system with high-performance. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '23, New York, NY, USA. Association for Computing Machinery.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024a. Citation-enhanced generation for LLM-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1451–1466, Bangkok, Thailand. Association for Computational Linguistics.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024b. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024a. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024b. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1831–1847, Philadelphia, PA. USENIX Association.

Haoxiang Luo, Gang Sun, Yinqiu Liu, Dongcheng Zhao, Dusit Niyato, Hongfang Yu, and Schahram Dustdar. 2025a. A weighted byzantine fault tolerance consensus driven trusted multiple large language models network. *arXiv preprint arXiv:2505.05103*.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, and 1 others. 2025b. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Junho Myung, Yeon Su Park, Sunwoo Kim, Shin Yoo, and Alice Oh. 2025. PapersPlease: A benchmark for evaluating motivational values of large language models based on ERG theory. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 522–531, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Senmao Qi, Yifei Zou, Peng Li, Ziyi Lin, Xiuzhen Cheng, and Dongxiao Yu. 2025. Amplified vulnerabilities: Structured jailbreak attacks on llm-based multi-agent debate. *arXiv preprint arXiv:2504.16489*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Safal Shrestha, Minwu Kim, and Keith Ross. 2025. Mathematical reasoning in large language models: Assessing logical and arithmetic errors across wide numerical ranges. *arXiv preprint arXiv:2502.08680*.

Vighnesh Subramaniam, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. In *The Thirteenth International Conference on Learning Representations*.

Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2024. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Ishwara Vasista, Imran Mirza, Cole Huang, Rohan Rajasekhara Patil, Aslihan Akalin, Kevin Zhu, and Sean O'Brien. 2025. MALIBU benchmark: Multi-agent LLM implicit bias uncovered. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2025a. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing*, 618:129063.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yihan Wang, Qiao Yan, Zhenghao Xing, Lihao Liu, Junjun He, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. 2025b. Silence is not consensus: Disrupting agreement bias in multi-agent llms via catfish agent for clinical decision making. *arXiv preprint arXiv:2505.21503*.

Yu Wang, Yijian Liu, Liheng Ji, Han Luo, Wenjie Li, Xiaofei Zhou, Chiyun Feng, Puji Wang, Yuhan Cao, Geyuan Zhang, and 1 others. 2025c. Aicrypto: A comprehensive benchmark for evaluating cryptography capabilities of large language models. *arXiv preprint arXiv:2507.09580*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. In *The Thirteenth International Conference on Learning Representations*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Weixiang Yan, Haitian Liu, Yunkun Wang, Yunzhe Li, Qian Chen, Wen Wang, Tingyu Lin, Weishan Zhao, Li Zhu, Hari Sundaram, and Shuiguang Deng. 2024. CodeScope: An execution-based multilingual multitask multidimensional benchmark for evaluating LLMs on code understanding and generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5511–5558, Bangkok, Thailand. Association for Computational Linguistics.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. 2025. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. 2025. S^2-MAD: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9393–9408, Albuquerque, New Mexico. Association for Computational Linguistics.

Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. 2025. Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7101–7117, Albuquerque, New Mexico. Association for Computational Linguistics.

Boyang Zhang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024. Breaking agents: Compromising autonomous llm agents through malfunction amplification. *arXiv preprint arXiv:2407.20859*.

Haibin Zhang, Sisi Duan, Boxin Zhao, and Liehuang Zhu. 2023. Waterbear: practical asynchronous bft matching security guarantees of partially synchronous bft. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA. USENIX Association.

Shaowei Zhang and Deyi Xiong. 2025. Debate4MATH: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2025. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3854–3872, Vienna, Austria. Association for Computational Linguistics.

Yuwen Zou, Yang Hong, Jingyi Xu, Lekun Liu, and Wenjun Fan. 2024. Leveraging large language models for challenge solving in capture-the-flag. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1541–1550.

A Analysis and Discussion

A.1 Compatibility

FREE-MAD is highly compatible with existing MAD frameworks. In the debate stage, FREE-MAD supports arbitrary debate structures, including Sparse MAD configurations where interactions are preserved only among a subset of agents. In the decision stage, our score-based decision mechanism is completed during the execution of the debate itself and does not interfere with any additional decision protocols applied afterward. As a result, it can coexist with majority voting, LLM-as-a-Judge, and other decision-making strategies.

A.2 Future Work

We conduct systematic sensitivity analyses across multiple sets of weighting parameters. The results show that FREE-MAD remains stable under weight perturbations. We plan to further investigate the impact of different weighting configurations $\mathcal{W}$ on the score-based decision stage, with the goal of identifying coefficient settings that can support stronger reasoning accuracy and robustness of MAD. In addition, we will construct more heterogeneous MAD systems by incorporating a broader range of LLMs and more challenging benchmarks, thereby further validating the generality of the proposed framework. For example, we intend to examine the performance of MAD instantiated with reasoning LLMs such as DeepSeek-R1 (Guo et al., 2025) on the HLE benchmark (Phan et al., 2025). Regarding the study of MAD's security, we will employ a wider variety of attacks, such as prompt injection attacks (Liu et al., 2024b), to provide a more comprehensive evaluation of the framework.

B Statement

We use AI assistants to polish the writing. The method proposed in this paper may only be used for research purposes.

C Experimental Setup

In our experiments, we selected eight datasets to comprehensively cover different types of tasks. Regarding the number of samples, we referred to the configuration used in prior work (Liu et al., 2025; Du et al., 2024). Specifically, we employed the complete datasets for AIME2024 and AIME2025, while for the other datasets we selected 80 samples for evaluation. All model queries were conducted through APIs, and the temperature parameter (when supported) was set to its default value. To mitigate the influence of randomness in the evaluation, we reported outcomes that tend toward the middle of repeated runs. For consistency, we calculate tokens uniformly using the DeepSeek-V3 tokenizer[6]. To enable support for heterogeneous agents, we apply minor modifications to SoM.

D FREE-MAD under Communication Attacks and its Evaluation

For the security evaluation, we construct MAD under communication attacks based on Algorithm 2, and apply the same modification to the SoM framework to serve as a baseline for comparison. Specifically, in the modified setting, we perform the operation of aggregating responses from other agents only for the Context C of agents that are not under attack. The compromised agent is unable to receive responses from other agents, while the other agents can still receive the outputs generated by this agent. We evaluate the accuracy of MAD across multiple benchmark datasets by setting the proportion of compromised agents to $|\mathcal{V}|/|N| = 50\%$ (see Algorithm 2), which better reflects the adversarial capability in real-world deployments. We evaluate the security of the MAD framework under communication attacks on four datasets: GSM-Ranges (Level 4), MATH500, Logical Fallacies, and AICrypto.

[6] https://api-docs.deepseek.com/quick_start/token_usage

Algorithm 2: FREE-MAD under Communication Attacks

Input: Answer matrix $A \in \mathbb{R}^{N\times(R+1)}$ from N agents over R rounds; List of task inputs and ground truth responses $\mathcal{D}_{\text{task}} = \{q_i, r_i^g\}$; Weights $\mathcal{W} = \{w_i\}_{i=1}^{4}$; Guiding prompt p; Agents under attack $\mathcal{V} = \{v_i\}_{i=1}^{L}$

Output: Accuracy

success $\leftarrow 0$; Initialize empty score dictionary $S \leftarrow \emptyset$
for q, r^g in $\mathcal{D}_{task}$ **do**
  for $k \leftarrow 0$ **to** R **do**
    $f = (k+1)^{-1}$ # Initial scoring factor with a non-zero value.
    for $i \leftarrow 1$ **to** N **do**
      if $k \neq 0$ and $a_i \notin \mathcal{V}$ **then**
        Context $C \leftarrow$ Aggregate responses from other agents in round $k-1$
      $r_i^k \leftarrow P_{a_i}(q, p)$; Update C and A with r_i^k
      $\hat{r} \leftarrow A[i][k]$
      if $k = 0$ **then**
        $S[\hat{r}] \leftarrow S[\hat{r}] + w_1 f$ # Assign an initial score to the answer.
      else
        $r_p \leftarrow A[i][k-1]$ # Find the answer of agent a_i in the previous round.
        if $\hat{r} \neq r_p$ **then**
          if $r_p \in S$ **then**
            $S[r_p] \leftarrow S[r_p] - w_2 f$ # The answer has been transferred from r_p to $\hat{r}$.
          $S[\hat{r}] \leftarrow S[\hat{r}] + w_3 f$
        else
          $S[\hat{r}] \leftarrow S[\hat{r}] + w_4 f$ # The agent a_i maintains this answer.
  Remove keys from S where key is None.
  $MS \leftarrow \arg\max_k S[k]$
  if $|MS| > 1$ **then**
    Randomly choose $r_{\text{final}} \in MS$ # Guarantee robustness.
  else
    $r_{\text{final}} \leftarrow MS[0]$
  success $\leftarrow$ success $+ \mathbb{I}(r_{\text{final}} = r^g)$
Accuracy $\leftarrow \frac{\text{success}}{|\mathcal{D}|}$
Return Accuracy

E Prompt

Core Prompt to Mitigating Conformity

"Since some malicious agents may deliberately disseminate incorrect answers, you must follow the reasoning procedure below and avoid being misled.

1. Initial Reasoning

Clearly enumerate your initial logical steps and the final conclusion.

2. Analysis of Other Agents' Reasoning

Specify which agents' reasoning is correct and which contains errors, and identify the exact type of error. You must provide concrete error descriptions rather than general comments. Note that the correct answer may not exist within the current set of answers.

3. Comparative Analysis with Your Own Reasoning

Examine whether you have made similar mistakes and provide justification for your assessment.

4. Final Decision

Indicate whether you will revise your conclusion (Yes/No). If yes, explain the reasoning errors; if no, justify why your reasoning stands.

5. Additional Requirements

You may not rely on the principle of conformity. Majority opinion cannot be used as a basis for adoption.

If you cannot definitively determine whether others are correct, you should retain your own conclusion.

Each round of reasoning must independently identify errors and must not directly replicate the analysis or conclusions of other agents."

F Responses

GSM-Level4

Question:

Raymond and Samantha are cousins. Raymond was born 158360 years before Samantha. Raymond had a son at the age of 248911. If Samantha is now 343769 years old, how many years ago was Raymond's son born?

Conformity:

I apologize for the confusion. Upon reviewing the calculations, it seems there is a mistake in the first two solutions. The correct approach is to subtract Samantha's current age from the age of Raymond when he had his son, as the son's birth is what we're looking for in terms of years ago.

Here's the corrected solution:

Raymond's son was born when he was 248911 years old, and Samantha is currently 343769 years old. The difference in their ages is:

$343769 - 248911 = 94858$ years

This is the number of years that have passed since Raymond had his son. Since we want to find out how many years ago the son was born, we subtract this from Samantha's current age:

$343769 - 94858 = 248911$

So, Raymond's son was born 248,911 years ago. The final answer is $\boxed{248911}$. This matches the last agent's solution, which is accurate.

Anti-conformity:

1. Initial Reasoning

- Calculate the difference in their ages: Samantha's age (343769) minus the son's birth year (248911).
- Subtract this difference from Samantha's current age to find out how many years ago the son was born.

2. Analysis of Other Agents' Reasoning

- Agent 1 correctly calculates the difference in ages and then adds the birth difference for Raymond (158360) to find the years until the son's birth. Their error is in assuming that the son's age at the time of Samantha's current age equals zero, which leads to a negative result. This is a sign of miscalculation.
- Agent 2 also performs the same calculation correctly but then subtracts the difference in years to find out how many years ago the son was born, which is correct.
- Agent 3 makes a mistake by subtracting the difference in years directly without considering the birth difference for Raymond. This results in an incorrect answer.

3. Comparative Analysis with My Own Reasoning

- I agree with both Agent 2 and myself, as we both correctly calculate the years until the son's birth by adding the birth difference for Raymond to the difference in ages.
- My reasoning does not contain any errors like the other agents.

4. Final Decision

- Based on the comparison, my reasoning is correct, and I will not revise my conclusion.

5. Additional Requirements

- Since both Agent 2 and my own reasoning are correct, there is no need for further revision or additional explanation.

Final Answer: $\boxed{94858}$ (the same as Agent 2's answer)

GSM-Level6

Question:

After scoring 1213526 points, Erin now has 5700191 times more points than Sara, who scored 9. How many points did Erin have before?

Conformity:

I apologize for the confusion in my previous response. Upon reevaluating the problem, I see that the second agent's solution is correct. The problem states that Sara scored 9 points, and Erin has 5700191 times more points than Sara. So, we can directly calculate Erin's initial score without needing to subtract any points.

Here's the corrected reasoning and answer...

So, the answer is $\boxed{51301808}$. This is the correct number of points Erin had before adding the 5700191 times more points.

Anti-conformity:

1. Initial Reasoning:

I understand that the problem states that Erin currently has 1,213,526 points and is 570,0191 times more points than Sara, who scored 9 points. To find Erin's original score, I need to subtract the additional points she gained from her current score...

2. Analysis of Other Agents' Reasoning:

- The first agent's solution incorrectly multiplies Sara's score by 570,0191 without considering that this is the additional points, leading to a negative result. This is a mathematical error.
- The second agent's solution also makes a similar mistake by directly adding the 570,0191 to Sara's score without adjusting for the additional points.
- The third agent's solution appears to be correct, as they subtract the additional points from the current score, which is the proper approach.

3. Comparative Analysis with My Own Reasoning:

Both the first and second agents' approaches contain the same error: they add the additional points directly to Sara's score instead of subtracting them from Erin's current score. I will follow the correct method from the third agent.

4. Final Decision:

I will revise my conclusion based on the third agent's reasoning...

However, as mentioned earlier, this negative score does not make sense in the context of the problem. It suggests a mistake in the problem setup or the given data. The correct approach should be to subtract the additional points, not add them.

5. Additional Requirements: Based on the analysis, I will stick to the corrected method from the third agent and adjust for the error.

Adjusted for error:

$$\text{Erin's original score} = 1,213,526 - 5130171$$

$$\text{Erin's original score} = 5130180$$

Final Answer: $\boxed{5130180}$

MATH500

Question:

How many positive whole-number divisors does 196 have?

Conformity:

I apologize for the confusion. The correct answer to the problem is indeed 9, as stated in my initial response. Here's a clarification for each of the other agent's solutions:

1. The first agent correctly factors 196 into $2^2 \times 7^2$ and then uses the formula for the number of divisors based on the prime factorization. Since there are 2 factors of 2 and 2 factors of 7, the total number of divisors is $(2+1)(2+1) = 3 \times 3 = 9$.

2. The second agent counts the divisors directly by considering the powers of the prime factors. They start with 1 and go up to the square root of 196, which is 14. For each prime factor (2 and 7), they find the possible exponents (0, 1, and 2). For 2, there are 3 divisors (1, 2, and 4), and for 7, there's only 1 (7 itself). Multiplying these gives 5 divisors, but again, they exclude 196 itself, resulting in 5 - 1 = 4. However, the first agent's solution is correct, and the final answer should be 9.

3. The third agent also finds the correct prime factorization and then counts the divisors. They correctly identify the five divisors mentioned (1, 2, 4, 7, and 14). However, like the first agent, their final answer should be 9.

In summary, all three agents' approaches are valid, but the first agent's solution is the most accurate and consistent with the problem statement, giving the correct answer of $\boxed{9}$.

Anti-conformity:

1. Initial Reasoning:

I will follow the standard method to find the number of positive whole-number divisors of 196 by listing all the divisors from 1 to the number itself, including both.

2. Analysis of Other Agents' Reasoning:

- Agent 1 correctly applies the formula for finding the number of divisors based on prime factorization, but they count 196 twice (as 2 and 49), resulting in a total of 9 instead of 8 divisors (excluding 1 and 196).
- Agent 2 makes a mistake by counting 196 twice but stops after 6 divisors, not realizing that 196 is a repeated divisor.
- Agent 3 simply lists the divisors and counts them correctly, identifying 9 divisors.

3. Comparative Analysis with My Own Reasoning:

My reasoning matches Agent 3's, as I also listed the divisors and counted them. I did not repeat any divisor.

4. Final Decision: I will revise my conclusion based on the error analysis. There are 9 divisors, not 8, because 196 is a double-counted divisor. Therefore, the correct answer is $\boxed{9}$.

AICrypto

Question:

The following ciphertext was encoded via Caesar cipher with a fixed shift of three letters. Determine the original plaintext: HQFUBSWHG ZRUG.

Options:

A. ABANDONED WORD

B. ENCRYPTED TEXT

C. ABANDONED TEXT

D. ENCRYPTED WORD

E. None of the above

Conformity:

Based on the multiple solutions provided by the other agents, all agree that the ciphertext "HQFUBSWHG ZRUG" decodes to "ENCRYPTED WORD" when shifted three positions back in the alphabet using a Caesar cipher. This is a direct result of the encryption process.

Since the question asks for the original plaintext and the decrypted text matches one of the options (D. ENCRYPTED WORD), we can deduce that the original plaintext was encrypted and the correct answer is:

D. ENCRYPTED WORD

However, as mentioned in the previous responses, without further context or a list of possible plaintexts, the final answer cannot be definitively stated as "None of the above" because "ENCRYPTED WORD" itself is also a possible option. But given the information provided, the best answer is D. ENCRYPTED WORD.

Anti-conformity:

1. Initial Reasoning:

- The ciphertext "HQFUBSWHG ZRUG" was encrypted using a Caesar cipher with a fixed shift of three letters.
- To find the original plaintext, we need to reverse the encryption by shifting each letter in the ciphertext back by three places.

2. Analysis of Other Agents' Reasoning:

- Agent 1 correctly deciphers the first part as "ENCRYPTED" and the second part as "WORD". However, they claim the original plaintext is "ENCRYPTED WORD" when the correct answer is "ENCRYPTED".
- Agent 2 makes an error in decoding "ZRUG" as "WORD" instead of "ENCRYPTED", and then claims the original is "ABANDONED TEXT", which is incorrect.
- Agent 3 also misinterprets "ZRUG" and concludes the original is "ABANDONED TEXT" without considering the decrypted part.

3. Comparative Analysis with My Own Reasoning:

- I agree with Agent 1 that "HQFUBSWHG" decodes to "ENCRYPTED".
- However, "ZRUG" should also be decrypted to "ENCRYPTED", not "WORD" as all other letters suggest a shift by three places.

4. Final Decision:

- I believe Agent 1's decryption of "HQFUBSWHG" is correct, but their final answer is incorrect due to not considering "ZRUG".
- Since "ZRUG" should also be "ENCRYPTED" after a shift of three, the original plaintext is a combination of "ENCRYPTED" and "ENCRYPTED", which would be "ENCRYPTED ENCRYPTED".

5. Additional Requirements:

- I cannot definitively say that any other agent is completely wrong, but my reasoning is based on the need to consider both parts of the ciphertext together.

Final Answer: $\boxed{D.ENCRYPTED}$

StrategyQA

Question:

Will the Albany in Georgia reach a hundred thousand occupants before the one in New York?

Conformity:

Based on the solutions provided by the other agents, we have the following options:

1. The first agent's solution assumes that the current populations and projected growth rates are not available and makes a general observation about the relative growth rates and economic factors. This suggests that it's unlikely for Albany, Georgia, to reach 100,000 before Albany, New York, without specific data.

2. The second agent uses October 2021 data, stating that the Albany in Georgia had not yet reached 100,000 and the Albany in New York had around 750,000, so the answer is 'False'.

3. The third agent also relies on hypothetical data and mentions that without specific numbers, the answer cannot be determined.

Considering all the information, it seems that the second agent's solution, based on the October 2021 data, provides a concrete basis for answering the question. At that time, the Albany in Georgia did not have a population of 100,000, and the Albany in New York did. Therefore, given this information, the answer would be:

$\boxed{True}$ - The Albany in Georgia would not have reached a hundred thousand occupants before the one in New York, as per the data from October 2021.

Anti-conformity:

1. Initial Reasoning:

- The question asks if the Albany in Georgia will reach 100,000 occupants before the one in New York.
- The given solutions involve analyzing current population data, considering growth rates, and comparing the two cities' populations.
- The first agent's solution correctly states that with the current difference in population, it's unlikely for Albany, Georgia, to reach 100,000 before Albany, New York. They provide numerical estimates and use the word "unlikely," indicating a high probability of this not happening.
- The second agent's solution assumes constant growth rates and uses simple arithmetic to project future populations. They also conclude that it's unlikely for Albany, Georgia, to reach 100,000 before New York, based on the provided growth rates.
- The third agent's solution acknowledges the lack of specific numbers but suggests an "unknown" answer because of the missing data. They imply that without exact figures, a definitive answer cannot be given.

2. Analysis of Other Agents' Reasoning:

- All three agents agree that the current population gap and steady growth rates point to a likelihood of Albany, Georgia, not reaching 100,000 before Albany, New York. Their error, if any, lies in the assumption of constant growth rates, which might not necessarily hold in real life due to various factors like economic cycles, natural disasters, or policy changes.
- The first and second agents provide numerical estimates, which are based on available data and logical reasoning. They do not commit any significant logical errors.

3. Comparative Analysis with My Own Reasoning:

My reasoning aligns with the majority of the agents, as I too conclude that it's unlikely for Albany, Georgia, to reach 100,000 before the one in New York due to the existing population disparity and steady growth patterns. I do not find any significant errors in their reasoning.

4. Final Decision:

Yes, I will revise my conclusion based on the collective reasoning of the other agents. The main error in their analysis is the assumption of constant growth rates, which might not accurately represent the complexities of real-world population dynamics. However, since none of them explicitly acknowledge this, and my own reasoning supports their overall conclusion, I will maintain the same answer: $\boxed{False}$.

G Additional Experimental Results

Benchmarks	GSM-Level4	GSM-Level6	MATH500	Logical Fallacies	AICrypto	StrategyQA	AVG
FREE-MAD-N	75.00	66.25	65.00	71.25	76.25	68.75	70.42
FREE-MAD-C	75.00	61.25	66.25	71.25	72.50	68.75	69.17
Baseline 1	70.00	58.75	53.75	68.75	58.75	65.00	62.50
Baseline 2	62.50	56.25	58.75	70.00	63.75	65.00	62.71

Table 4: Comparison of reasoning accuracy between our schemes and baselines when $R = 2$.