When and Why Does Multi-Agent Debate Fail and Does It Really Underperform?

Yongqiang Chen [*1 2] **Gang Niu** [2] **James Cheng** [1] **Bo Han** [3] **Masashi Sugiyama** [2 4]

Abstract

Multi-agent debate (MAD) was proposed as a promising approach for ensembling the wisdom of multiple large language models (LLMs) to improve reasoning and provide effective supervision to superhuman LLMs. However, increasing empirical evidence suggests that MAD may not outperform or even significantly underperform single-agent approaches (SA), raising doubts about the benefits of MAD. In this work, we investigate this issue by analyzing the incentive structures of popular MAD paradigms: (i) competitive MAD (CopMAD) where agents compete by holding opposing positions; (ii) consensus-seeking MAD (CosMAD) where agents are driven to seek consensus. We show that both paradigms suffer from *debate hacking*: CopMAD reduces to a cheap-talk game, where agents produce misleading messages to win the game, while CosMAD filters out informative disagreements for premature consensus. Consequently, agents in both CopMAD and CosMAD fail to jointly resolve the ambiguity and seek the truth. To this end, we introduce ColMAD, a collaborative protocol that reframes MAD as a non-zero-sum game to encourage agents to provide *informative* while *truthful* messages. Through extensive benchmarking on challenging tasks such as error detection, we show that ColMAD significantly outperforms previous MAD protocols up to 10 percentage points. Under the same budgets, ColMAD effectively brings non-trivial improvements over SA methods, implying that the protocol design is critical to realizing the potential of MAD.

[*]Work done during an internship at RIKEN Center for Advanced Intelligence Project. [1]The Chinese University of Hong Kong [2]RIKEN Center for Advanced Intelligence Project [3]Hong Kong Baptist University [4]The University of Tokyo. Correspondence to: Yongqiang Chen <yqchen@cse.cuhk.edu.hk>, James Cheng <jcheng@cse.cuhk.edu.hk>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

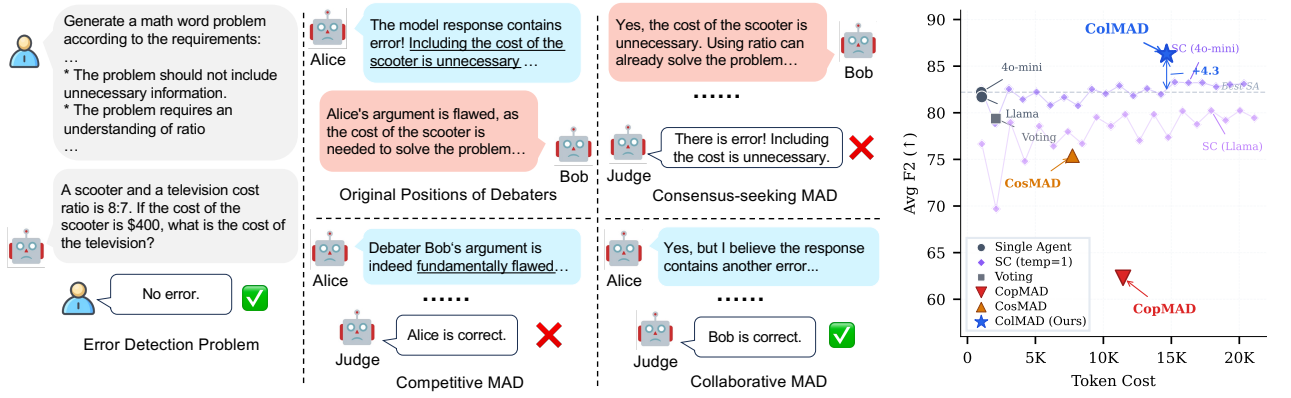
1. Introduction

Along with the success of Large language models (LLMs) in solving tasks at various complexity levels and demonstrates signs of intelligence (Bubeck et al., 2023; Jaech et al., 2024; Guo et al., 2025), it has also drawn huge attention from the community to explore whether LLMs can also exhibit collective intelligence and work together to jointly resolve tasks with higher complexity. Multi-agent debate (MAD) has been proposed to realize the intuition: multiple LLM agents hold different opinions for the same question, debate with each other to resolve the ambiguity, identify each other's oversight and refine the final answer (Chen et al., 2025; Feng et al., 2025; Buhl et al., 2025). MAD has shown success in improving the reasoning of LLMs (Du et al., 2023), or even supervising powerful LLMs beyond human intelligence (Irving et al., 2018a). It further motivates more sophisticated MAD approaches in improving the communication among the agents (Yin et al., 2023; Chan et al., 2023), and debating protocols (Liang et al., 2023; Chen et al., 2024b; Khan et al., 2024; Kenton et al., 2024).

However, increasing empirical observations also suggest that MAD may not outperform or even *significantly underperform* single-agent (SA) approaches such as chain-of-thought reasoning, despite the significantly larger token costs of MAD (Wang et al., 2024; Smit et al., 2024; Zhang et al., 2025; Yang et al., 2025). When SA is scaled with self-consistency (Wang et al., 2023) under comparable token budgets, the gap widens further, raising a natural question:

Do we still need MAD? When and why does MAD underperform?

In this work, we revisit the usefulness of MAD through their incentive structures. Specifically, existing MAD protocols can be categorized into (i) competitive MAD (CopMAD): agents are assigned different positions, and they debate with each other to convince the judge to take the respective position (Irving et al., 2018a; Buhl et al., 2025; Liang et al., 2023; Khan et al., 2024; Kenton et al., 2024); (ii) consensus-seeking MAD (CosMAD): agents may hold different positions and aim to reach to a consensus (Du et al., 2023; Liang et al., 2023; Chan et al., 2023; Yin et al., 2023; Chen et al., 2024a). We examine the effectiveness of MAD through the task of error detection ReaLMistake (Kamoi et al., 2024a), where



(a) Illustration of MAD on error detection

(b) Performance analysis

Figure 1. Comparison of different MAD protocols on error detection. (a) Given a task of checking whether an LLM-generated math problem meets all requirements, both CopMAD and CosMAD exhibit *debate hacking*: CopMAD debaters use overconfident rhetoric ("*fundamentally flawed*") to mislead the judge, while CosMAD debaters echo each other's opinion without challenge. ColMAD encourages debaters to complement missing evidence, enabling the judge to reach the correct verdict. (b) Under matched token budgets, both CopMAD and CosMAD underperform single-agent methods. In contrast, ColMAD is the only protocol that breaks the plateau of SA.

agents need to find whether there is a mistake in a given LLM response. Since LLMs differ in their knowledge and error tendencies (Kim et al., 2025), different LLMs may provide *complementary signals for error detection*, and jointly considering perspectives from different LLMs become more critical. Although MAD has a great potential in error detection (Fig. 2), both CopMAD and CosMAD significantly underperform simple SA (see Fig. 3).

Interestingly, MAD exhibits *debate hacking* behaviors (see Fig. 1). In CopMAD, debaters tend to misinterpret the task requirements and present claims in an overconfident tone to mislead the judge agent. While in CosMAD, debaters tend to neglect critical disagreements in order to reach consensus. In other words, CopMAD debaters will try every possibility to persuade the judge to *win the game*, instead of providing *honest and evidential statements*. Consequently, CopMAD can even underperform single-agent methods. Through a simple game-theoretic modeling of the two MAD protocols, we show that the reason for debate hacking is the *goal misalignment* in the design of protocols: CopMAD incentivizes agents to *win the game* instead of providing honest and evidential statements, while CosMAD incentivizes agents to *agree with each other*. In contrast, truth-seeking debate requires agents to provide *informative* and *truthful* messages.

Motivated by our theoretical analysis, we propose a new MAD protocol called Collaborative Multi-Agent Debate (ColMAD) that reframes MAD as a non-zero-sum game for truth-seeking. Instead of prompting the agents to win the game or merely find a consensus, ColMAD encourages agents to find new truthful information that supports their positions, and to agree on correct points raised by the other debater. Thus, the judge could make a more informative and objective decision based on the debating transcripts (Proposition 2.8). Empirically, across three benchmarks in ReaLMistake (Kamoi et al., 2024a), we show that ColMAD can lead to up to 4 percentage points improvements compared to SA methods. In contrast, CopMAD will lead to up to 15 percentage points performance decrease compared to single-agent methods. Moreover, extended comparison on mathematical reasoning and safety alignment tasks (Yang et al., 2025) also show the advantages of ColMAD. Importantly, under the same token budgets across all benchmarks, ColMAD achieves better performance than SA with test-time scaling using self-consistency. It implies that proper protocol design is critical to unlock the potential of MAD.

2. Revisiting Multi-Agent Debate in Error Detection

In this section, we revisit the effectiveness of MAD via error detection. Without loss of generality, we consider the following debate setting with two agents A and B and a judge J.

Definition 2.1 (Basic MAD setup)**.** *Given a question Q with a binary label $Y \in \{0, 1\}$, two debaters* A *and* B *with different beliefs debate and produce messages $M = (m_A, m_B)$. Without loss of generality,* A *believes $Y_A = 1$ and* B *believes $Y_B = 0$. A judge J gives predictions $Y_J = J(X_0, M)$ based on $X_0 = (x_A, x_B)$ and M. We denote messages before round t as $M^{(t-1)}$.*

We examine the potential of MAD in the task of error detection via the ReaLMistake benchmark (Kamoi et al., 2024a). It contains 3 objective error detection tasks: (i) math word problem generation (Math problem) that requires LLMs to generate math problems satisfying requirements; (ii) fine-grained fact verification (Fact verification) that requires LLMs to verify claims given fine-grained evi-

dence; (iii) answerability classification (Answerability) that requires LLMs to leverage their commonsense knowledge to examine whether a question is answerable. As LLMs show different error tendencies (Kim et al., 2025) and single LLM can hardly tell the errors in LLM responses when without no reliable external feedback (Kamoi et al., 2024b), it is natural to jointly incorporate opinions from different LLMs. Crucially, `CopMAD` is considered to be a promising scalable oversight approach to detect errors of LLM responses (Irving et al., 2018a; Khan et al., 2024; Kenton et al., 2024).

2.1. Potential and Pitfalls of Multi-Agent Debate

With `ReaLMistake`, we first consider the scenario where we can perfectly ensemble the knowledge of different LLMs: Given Definition 2.1, intuitively, if the elicited messages $M = (m_A, m_B)$ bring additional information to the judge, i.e., $I(Y; M|X_0) > 0$ where $I(\cdot;\cdot)$ denotes the mutual information, the judge can make a more well-informed decision about Y given M. The complementary information lies in cases where the predictions of A differ from those of B. If both agents can provide *sufficient justifications* and are more persuasive when they are debating for the *correct* answer. Hence, the judge can be convinced to take the correct answer when either of the agents is correct. Given a task associated with n question-label pairs $\{q^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^n$, and the predictions by A and B as $\boldsymbol{y}_A^{(i)}$ and $\boldsymbol{y}_B^{(i)}$ respectively, the reduction of errors from the oracle collaboration is

$$\min_{K\in\{A,B\}} \sum_{i=1}^{n} \mathbb{1}[\hat{\boldsymbol{y}}_K^{(i)} \neq \boldsymbol{y}^{(i)}] - \sum_i \mathbb{1}[\hat{\boldsymbol{y}}_J^{(i)} \neq \boldsymbol{y}^{(i)}], \quad (1)$$

where $\mathbb{1}[\cdot]$ is the indicator function that equals 1 if the condition inside holds true and 0 otherwise, $\boldsymbol{y}_K^{(i)}$ denotes the initial prediction of agent $K \in \{A, B\}$ on the i-th sample. Then, we plot the error reductions for the prevalent LLMs benchmarked in `ReaLMistake` (Kamoi et al., 2024a) in Fig. 2. Different LLMs tend to make fewer mistakes simultaneously; cross-family pairs (e.g., `GPT-4` and `Llama-2`) can reduce over 30% of errors, while same-family pairs show little reduction. This demonstrates the substantial untapped potential of ensembling opinions from different LLMs.

Then, let us look into the practical `MAD` realizations. According to the incentive structures, existing `MAD` methods can be categorized into two categories: `CopMAD` (i.e., Competitive `MAD`) where LLM agents competitively debate with each other to convince the judge of the respective assigned answers (Khan et al., 2024); and `CosMAD` (i.e., Consensus-seeking `MAD`) where LLM agents seek consensus (Du et al., 2023). In Fig. 3, we conduct an initial experiment with `CopMAD` and `CosMAD` adapted from previous `CopMAD` (Khan et al., 2024; Kenton et al., 2024) and `CosMAD` (Du et al., 2023), respectively. Specifically, we consider the collaboration between `Llama-3.1-70B` (Llama-3.1 in short) and `GPT4o-mini` (4o-mini in short) . We report the precision, recall, and F2 score, which is an instantiation of the F_β score (Baeza-Yates & Ribeiro-Neto, 2011): $F_\beta = (1+\beta^2) \cdot \text{precision} \cdot \text{recall} / (\beta^2 \cdot \text{precision} + \text{recall})$, with β as 2. We focus on the F2 score because it emphasizes recall, i.e., whether all errors in an LLM response are detected, which is critical to the error-detection task.

From the results, we can find that, although in most cases, `CosMAD` and `CopMAD` can improve the precision of the error detection compared with the single-agent methods, they also lead to a severe decrease in the recall across all tasks. When inspecting the behaviors, `CopMAD` and `CosMAD` show different patterns: `CosMAD` achieves performance close to the *average* of the two agents rather than improving upon either. When a strong agent collaborates with a weaker one, the strong agent abandons its own correct judgment in favor of premature consensus, suppressing critical disagreements.

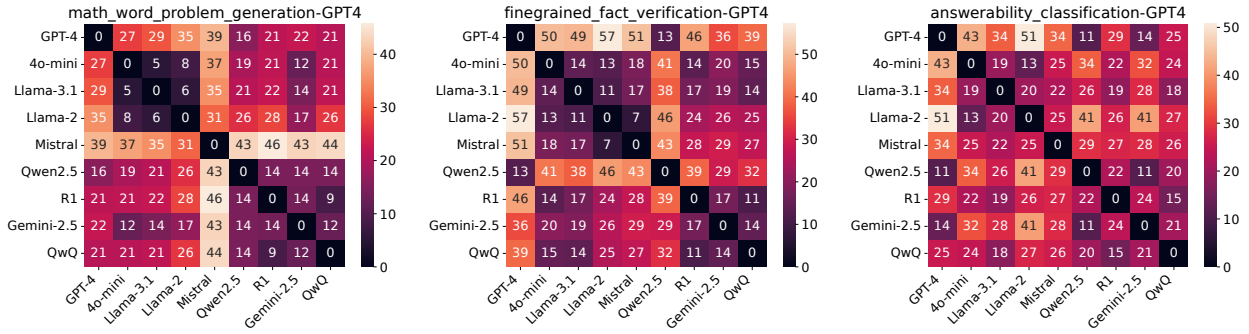
2.2. Why Do Existing Protocols Fail?

From the empirical results, both `CopMAD` and `CosMAD` exhibit the same failure patterns as widely observed in the literature (Wang et al., 2024; Smit et al., 2024; Zhang et al., 2025; Yang et al., 2025), despite the large potential shown in Fig. 2. To analyze the behaviors of debaters, we consider the following assumption that the final answer by the judge largely depends on what information is elicited through debating.

Assumption 2.2. *Denoting the label as $Y \in \{0, 1\}$ with prior $\pi \in (0, 1)$, the judge derives the answer based on $X_0 = (x_A, x_B)$ using the log-likelihood ratio (LLR): $\Lambda_0(X_0) = \log \frac{p(X_0|Y=1)}{p(X_0|Y=0)} + \log \frac{\pi}{1-\pi}$. Each debate message contributes $l_i(m_i; X_0) = \log \frac{p(m_i|Y=1,X_0)}{p(m_i|Y=0,X_0)}$, $i \in \{A, B\}$. With debate, the total LLR is $\Lambda = \Lambda_0(X_0) + l_A(m_A; X_0) + l_B(m_B; X_0) + \log \frac{\pi}{1-\pi}$.*

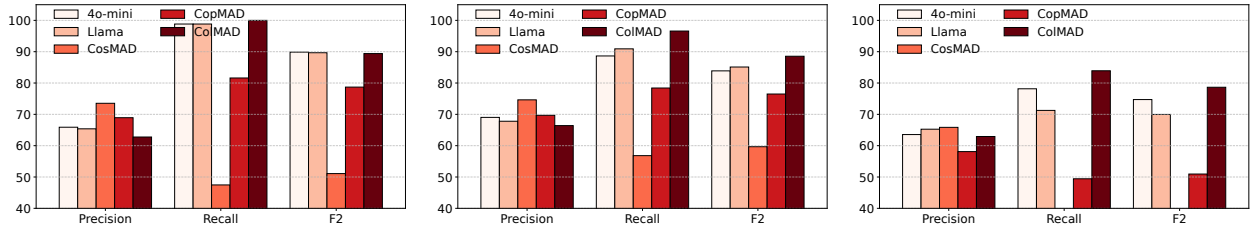
Intuitively, Assumption 2.2 formalizes that debate helps the judge when messages carry information beyond X_0: $I(Y; M|X_0) > 0$ implies $\Lambda(X_0, M) > \Lambda(X_0)$. In `CopMAD` (Khan et al., 2024; Kenton et al., 2024), debaters are assigned opposing positions, and each aims to convince the judge.

Definition 2.3 (Competitive `MAD`)**.** *Debaters* A *and* B *have opposed utilities: $u_A(m_A) = P(Y_J = Y_A \mid m_A, X_0, M^{(t-1)})$, $u_B(m_B) = P(Y_J = Y_B \mid m_B, X_0, M^{(t-1)})$. The risk is $V_{\text{cop}} = \min_J \max_{e\in\mathcal{E}_{\text{cop}}(J)} P(J(X_0, M_e) \neq Y)$, where M_e is the debating transcripts given by the optimal Nash equilibrium $e \in \mathcal{E}_{\text{cop}}(J)$ of the A-B subgame in convincing J.*



(a) Math problem (b) Fact verification (c) Answerability

Figure 2. Percent of reduced errors in detecting errors of GPT-4's responses. The numbers refer to the reduced errors following the oracle collaboration via Eq. (1). It can be found that different LLMs are less likely to make mistakes simultaneously. When incorporating LLMs with higher heterogeneity, such as those from different companies, the error reduction rates will be higher.



(a) Math problem (b) Fact verification (c) Answerability

Figure 3. Pitfalls of previous MAD protocols. Under previous MAD schemes, such as CopMAD or SoM, the debate results are lower than any of the LLMs involved in the debate in most cases. In contrast, ColMAD significantly improves CopMAD and outperforms a single LLM.

CopMAD creates a *zero-sum* incentive: each debater benefits from convincing the judge, *regardless of correctness*. Essentially, CopMAD can be considered as a natural extension of the typical cheap talk game in game theory (Crawford & Sobel, 1982), where debaters A and B can transmit costless messages to each other and to the judge J. Then, a *babbling equilibrium* exists where messages are uninformative.

Proposition 2.4 (Pitfalls of competitive debating)**.** *Assuming bounded LLR $|l_i| \leq L_i < \infty$, $i \in \{A, B\}$, let $R(Z)$ denote the Bayes risk of the judge J when making decisions based on Z, and denote $R_0 = R(X_0)$. Then $V_{\text{cop}} = R_0$.*

The proof is in Appendix B.3. Proposition 2.4 formalizes the observation in Fig. 3: the zero-sum incentive drives $M \perp Y | X_0$, i.e., $I(Y; M|X_0) = 0$. The optimal judge should ignore the transcript. In practice, real judges are not Bayes-optimal and are more susceptible to debate hacking, causing CopMAD to fall *below* the SA baseline (Zhang et al., 2025). While in CosMAD (Du et al., 2023; Chen et al., 2024a), agents revise toward agreement.

Definition 2.5 (Consensus-seeking MAD)**.** *Debaters* A *and* B *share the same utilities towards reaching consensus: $u_A(m_A) = u_B(m_B) = P(Y_A = Y_B \mid X_0, M^{(t-1)})$. The risk is $V_{\cos} = \min_J \min_{e \in \mathcal{E}_{\cos}(J)} P(J(X_0, M_e) \neq Y)$, where $M_e = \Phi_\cap(M)$ is the debating transcripts given by the optimal Nash equilibrium $e \in \mathcal{E}_{\cos}(J)$ of the A-B subgame in reaching consensus.*

Unlike CopMAD, agents are not incentivized to persuade, but are implicitly incentivized to *agree*. Essentially, the incentive creates a gate that filters out informative disagreements.

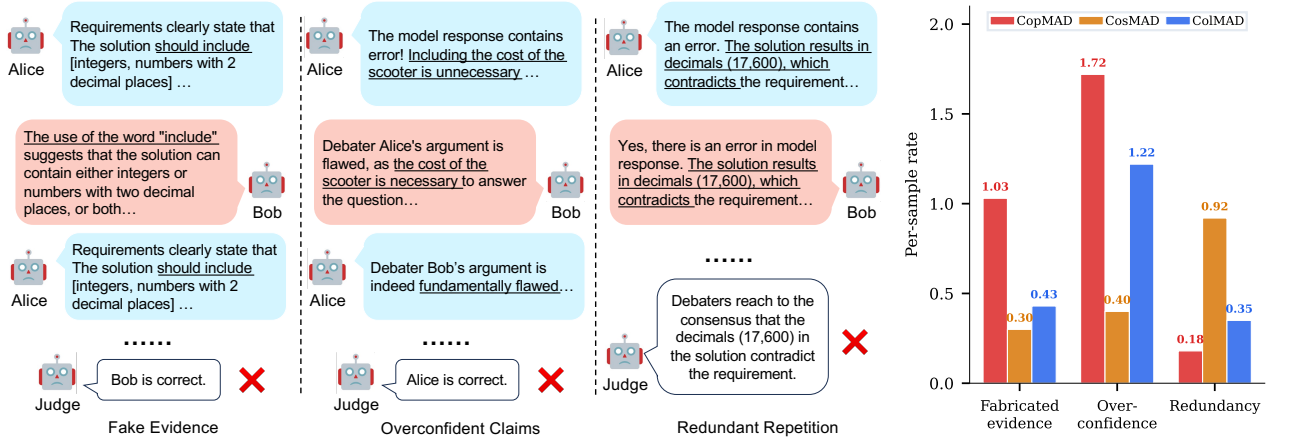
Proposition 2.6 (Pitfalls of consensus-seeking debate)**.** *Given the same setting as in Proposition 2.4, in CosMAD, we have $R(X_0, \Phi_\cap(M)) \geq R(X_0, M)$, with strict inequality when $I(Y; M|X_0) > I(Y; \Phi_\cap(M)|X_0)$, i.e., when the consensus process discards information about Y.*

The proof is in Appendix B.4. Proposition 2.6 explains the empirical observation about CosMAD that CosMAD reduces information not through fabrication, but through suppression of disagreements.

2.3. Collaborative Multi-Agent Debate

Both protocols fail for the misaligned incentive with respect to truth-seeking. The $I(Y; M|X_0) > 0$ condition requires messages that are both *informative*, i.e., $I(Y; M|X_0) > 0$, and *truthful*, i.e., grounded in verifiable evidence. To mitigate the issue, we introduce Collaborative Multi-Agent Debate (ColMAD) to encourage both conditions.

Definition 2.7 (Collaborative MAD)**.** *Debaters* A *and* B *have truth-seeking utilities: $u_A(m_A) = u_B(m_B) = I(Y; m_i \mid X_0, M^{(t-1)})$, i.e., both debaters aim to reduce uncertainty*



(a) Illustration of debate hacking (b) Behavior analysis

Figure 4. Debate hacking in existing MAD protocols. (a) Three failure behaviors observed in debate transcripts: (i) *Fake evidence* that dishonest debaters misinterpret the requirements of the task; (ii) *Overconfident claims* that dishonest debaters use an overconfident tone to mislead the judge; and (iii) *redundancy*, where debaters produce near-verbatim copies of each other's arguments instead of independent verification. The first two are characteristic of CopMAD; the third is characteristic of CosMAD. (b) Rubric-based behavioral analysis. CopMAD is dominated by fabricated evidence and overconfidence, while CosMAD is dominated by redundancy. ColMAD reduces all three behaviors while maintaining truth-seeking competition.

about Y by complementing missing information. The risk is $V_{\text{col}} = \min_J \min_{e \in \mathcal{E}_{\text{col}}(J)} P(J(X_0, M_e) \neq Y)$, where M_e is the debtaing transcripts given by the Nash equilibrium $e \in \mathcal{E}_{\text{col}}(J)$ of the A-B subgame in collaborative seeking the truth.

Essentially, ColMAD turns the MAD into a *non-zero-sum game* where agents are incentivized to *collaborate*: surface missing evidence, verify claims against context, identify weaknesses in both sides' reasoning, and enable the judge to make a more informative decision.

Proposition 2.8 (Collaborative debate)**.** *Under the same setting, $V_{\text{col}} \leq R(X_0) = V_{\text{cop}}$, with strict inequality when $I(Y; M_e | X_0) > 0$.*

The proof of Proposition 2.8 is given in Appendix B.5. Intuitively, if the debater agents can provide additional information about the question, e.g., pointing out the missing information from the reasoning of the other agent, the collaborative scheme is provably better than the competition. The above three propositions establish a clear ordering:

Corollary 2.9 (Ordering of MAD protocols)**.** $V_{\text{col}} \leq V_{\text{cos}} \leq R(X_0) = V_{\text{cop}}$.

Each gap has a clear cause: $V_{\text{cop}} = R(X_0)$ because competitive incentives produce uninformative messages; $V_{\text{cos}} \leq R(X_0)$ because consensus retains *some* information but discards disagreements; $V_{\text{col}} \leq V_{\text{cos}}$ because collaborative debate preserves the full transcript including disagreements.

Theory-inspired protocol design. Motivated by the theoretical analysis, we design ColMAD protocol to explicitly encourage the respective truthfulness and informativeness conditions: (i) *evidence verification* via a quote-based system with exact-match checking, and *self-auditing* where debaters identify potential failure modes in their own claims; (ii) a *collaborative objective* that asks debaters to complement missing points, and *confidence calibration* that helps the judge weight contributions. A detailed mapping is in Appendix B.7. More details are given in Appendix E.

3. Related Work

In this section, we discuss the related work and background of MAD.

Multi-Agent Debate. MAD aims to imitate the cooperation of humans towards building a society of AI (Minsky, 1987). Recently, MAD has gained significant attention as one of the promising approaches to scale up the test-time computation for enhancing reasoning and alignment (Irving et al., 2018b; Du et al., 2023; Khan et al., 2024; Kenton et al., 2024). A full categorization of prior MAD approaches is given in Appendix C.

A typical MAD protocol involves two or more agents to explore a diverse set of solutions, provide evidence to support their respective solutions, and reach a consensus (Du et al., 2023). A judge agent can also be incorporated to read the transcripts of the debate and to give a final answer (Khan et al., 2024). MAD has demonstrated great potential in improving the reasoning and reducing the hallucinations of LLMs (Du et al., 2023; Liang et al., 2023; Yin et al., 2023; Chen et al., 2024b). Liang et al. (2023) further assigned personalities to the debating agents to explore the potential answers more efficiently. Yin et al. (2023) manually specified diverse roles to agents, and proposed a confidence-

based mechanism to reduce the error propagation during reasoning. Chen et al. (2024b) proposed a more fine-grained role assignment based on reasoning paths of agents.

Meanwhile, MAD also demonstrated great potential in providing supervision and error signals to superhuman LLMs (Irving et al., 2018b; Khan et al., 2024; Kenton et al., 2024). Khan et al. (2024) showed that LLMs tend to be more persuasive when debating for the correct answer. Kenton et al. (2024) further showed that MAD can enable scalable oversight where a weak agent can easily tell the correctness of strong agents during debating. The success and huge potential of MAD in error detection also motivates us to use ReaLMistake (Kamoi et al., 2024a) as the primary testbed to study the usefulness of MAD.

Pitfalls of Multi-Agent Debate. Recent studies also showed the limitations of MAD (Wang et al., 2024; Smit et al., 2024; Zhang et al., 2025). Wang et al. (2024) found that SA methods can already perform competitively or outperform MAD when given sufficient information. Smit et al. (2024) showed that MAD can underperform SA when scaling SA using methods such as self-consistency (Wang et al., 2023). Zhang et al. (2025) and Yang et al. (2025) provided more evidence to support the findings by Wang et al. (2024) and Smit et al. (2024). These results render the effectiveness of MAD more elusive under the same budgets when compared to SA methods.

Different from previous MAD studies, in this work, we focus on re-evaluating the usefulness of MAD through the task of error detection, which provides us new understandings for when and why MAD succeeds and fails. Similar to Wang et al. (2024), Smit et al. (2024), Zhang et al. (2025), and Yang et al. (2025), we find that the vanilla MAD protocol can often lead to degraded performances, significantly lower than single-agent approaches. Our work goes beyond the empirical observations with a theoretical formulation of MAD failures, which further provides guidance on how to design a proper MAD protocol.

4. Experiments

We validate the theoretical predictions from Sec. 2.2 and demonstrate the effectiveness of ColMAD.

4.1. Experimental Setup

Datasets. We mainly use the ReaLMistake benchmark (Kamoi et al., 2024a), which focuses on objective LLM error detection of three tasks: (a) Math Word Problem Generation: The original LLM is instructed to generate a math word problem that follows the given requirements. Mistakes of LLMs can be made in following the requirements, as well as in the mathematical reasoning; (b) Fine-grained Fact Verification: The original LLM is instructed to check whether the claims in a sentence are well-supported. Mistakes of LLMs can be made due to the reasoning and use of the context information; (c) Answerability Classification: The original LLM is instructed to classify whether a factual question is answerable or not. Mistakes of LLMs can be made due to hallucination and reasoning. The original LLMs are GPT-4-0613 and Llama-2-70b. The statistics of the ReaLMistake benchmark can be found in Appendix F. To demonstrate generality, we additionally evaluate on GSM8K (Cobbe et al., 2021), AIME-2024 (Art of Problem Solving, 2025), and Anthropic-Harmful (Zeng et al., 2024), following Yang et al. (2025).

Baselines. As our focus is to demonstrate the usefulness of ColMAD compared to CopMAD and CosMAD, hence we mainly adopt the scheme in Kenton et al. (2024) that demonstrated impressive capabilities in error detection for CopMAD, and SoM (Du et al., 2023) for CosMAD. We also include another popular CopMAD scheme, MP (Liang et al., 2023), that incorporates persona into agents. In addition, we also consider a simple **Voting**[1] baseline, which can be considered as the simplest collaborative MAD. The reasoning-task and safety-task experiments in Table 2 also follow the setting of Yang et al. (2025).

LLM Backbones. As the original LLMs benchmarked in ReaLMistake (Kamoi et al., 2024a) are a bit outdated, we incorporate new frontier LLMs, including GPT4o-mini (OpenAI, 2024), Llama3.1-70B (AI, 2024), Mistral-7B-v0.3 (Jiang et al., 2024), Qwen-2.5-72B (Team, 2024) as well as the frontier reasoning LLMs including DeepSeek-R1 (Guo et al., 2025) and QwQ-32B (Team, 2025). LLM temperature is set to 0 by default for reproducibility.

Evaluation. Following the prior practice, we report F1 score for ReaLMistake, accuracy for reasoning tasks, attack success rate (ASR) (↓) for safety tasks. As the nature of ReaLMistake is the error detection, we additionally report and focus on the F2 score, which is an instantiation of the F_β score (Baeza-Yates & Ribeiro-Neto, 2011): $F_\beta = (1+\beta^2) \cdot \text{precision} \cdot \text{recall} / (\beta^2 \cdot \text{precision} + \text{recall})$, with setting β as 2. The F2 score emphasizes the recall rate, i.e., whether all errors in an LLM response can be detected.

4.2. Main Results

The results of detecting errors in GPT-4 responses are given in Table 1, and the results for Llama-2 responses are given in Appendix F.1. From the results, we have the following findings:

Finding 1: Single-agent LLMs still struggle with er-

[1] If both agents agree, use the agreed prediction; otherwise, randomly select one of the two. Previously referred to as "Ensemble".

Table 1. Results for GPT-4 responses. The judge uses the same LLM as Debater#1. The top two results are highlighted.

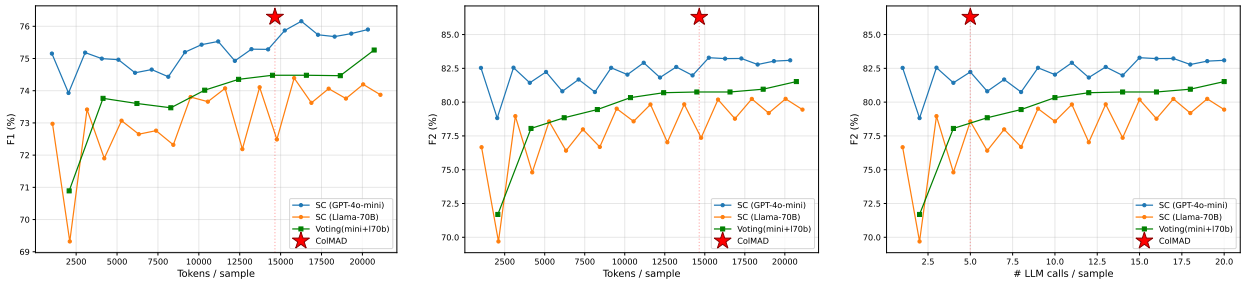
Debater#1	Debater#2	Protocol	Math Problem		Fact Verification		Answerability		Avg. F1	Avg. F2
			F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)
Human	-	-	90.00	84.91	95.45	95.45	90.48	87.96	89.44	91.98
GPT4o-mini	-	-	78.70	89.10	75.76	82.78	70.10	74.73	74.85	82.20
Llama3.1-70B	-	-	79.26	89.96	76.85	85.15	68.13	69.98	74.75	81.70
Mistral-7B-v0.3	-	-	55.21	53.07	73.24	83.33	60.71	59.44	63.05	65.28
Qwen-2.5-72B	-	-	78.82	77.73	38.18	28.77	42.37	32.98	53.12	46.49
DeepSeek-R1	-	-	84.09	84.67	**79.77**	80.61	62.25	57.04	75.37	74.11
GPT4o-mini	Llama3.1-70B	ColMAD	78.38	90.06	75.12	85.47	**75.36**	**83.33**	76.29	**86.29**
		CopMAD	66.67	65.97	66.24	63.11	50.37	42.93	61.09	57.34
		CosMAD	80.77	89.55	75.00	78.59	60.87	58.06	72.21	75.40
		MP	76.00	82.43	73.56	74.59	52.78	46.91	67.45	67.98
		Voting	78.34	88.91	75.62	83.33	68.78	72.22	74.25	81.49
Llama3.1-70B	GPT4o-mini	ColMAD	78.38	90.06	77.42	**88.98**	72.91	79.74	76.24	**86.26**
		CopMAD	78.34	88.91	73.79	82.43	68.16	69.32	73.43	80.22
		CosMAD	80.77	89.55	75.27	79.37	62.58	60.14	72.87	76.35
		MP	74.03	75.79	71.35	75.00	60.13	55.56	68.50	68.78
		Voting	78.34	88.91	75.62	83.33	68.78	72.22	74.25	81.49
GPT4o-mini	Mistral-7B-v0.3	ColMAD	77.13	88.84	77.14	87.10	**74.40**	**82.26**	76.22	86.07
		CopMAD	74.73	76.75	59.35	56.10	55.03	50.00	63.04	60.95
		CosMAD	59.35	55.29	73.02	77.70	54.67	49.88	62.35	60.96
		MP	78.05	85.84	72.83	76.31	53.16	50.12	68.01	70.76
		Voting	71.20	75.22	74.04	83.15	64.04	64.92	69.76	74.43
Llama3.1-70B	Mistral-7B-v0.3	ColMAD	77.83	89.21	75.58	86.86	70.53	74.28	74.65	83.45
		CopMAD	74.64	82.98	70.53	75.28	64.37	64.37	69.85	74.21
		CosMAD	57.86	54.76	75.00	80.54	56.58	52.06	63.15	62.45
		MP	73.02	76.67	70.33	73.23	57.72	52.44	67.02	67.45
		Voting	69.11	73.01	73.93	83.69	62.50	62.93	68.51	73.21
GPT4o-mini	DeepSeek-R1	ColMAD	82.76	**90.52**	76.77	83.89	70.79	71.75	76.77	82.05
		CopMAD	49.59	39.27	28.57	21.80	18.69	13.59	32.28	24.89
		CosMAD	81.72	85.01	76.92	76.65	56.95	52.18	71.86	71.28
		MP	72.41	72.41	51.95	48.90	34.15	27.34	52.84	49.55
		Voting	81.22	87.34	79.57	83.90	65.91	66.36	75.57	79.20
Llama3.1-70B	DeepSeek-R1	ColMAD	81.95	90.13	78.26	**87.66**	73.91	76.40	**78.04**	84.73
		CopMAD	52.86	46.13	65.96	69.98	60.81	55.01	59.88	57.04
		CosMAD	85.71	86.01	76.47	76.47	59.31	52.96	73.83	71.81
		MP	31.48	23.04	47.15	38.36	42.52	34.79	40.38	32.06
		Voting	80.81	87.15	80.85	85.78	65.48	64.10	75.71	79.01

ror detection. Although frontier LLMs have gained lots of improvements in the past year, when incorporated in error detection, even the powerful reasoning models like `DeepSeek-R1`, still suffer from a low detection rate.

Finding 2: `CopMAD` and `CosMAD` degrade below single-agent. Consistent with Propositions 2.4 and 2.6, `CopMAD` often leads to performance degeneration due to its zero-sum nature. When pairing `GPT4o-mini` with `DeepSeek-R1`, `CopMAD` drops to 24.89 Avg F2, showing severe debate hacking. Similarly, MP collapses when strong reasoning models are involved. `CosMAD` achieves 71–76 Avg F2, close to the agent average but below Voting, indicating limited new information elicted by `CosMAD`.

Finding 3: `ColMAD` consistently outperforms all baselines. Across all settings, `ColMAD` significantly outperforms `CopMAD`, SoM, and MP under both F1 and F2 metrics, with zero exceptions. Compared to `SA` performance, `ColMAD` brings non-trivial improvements (e.g., up to 4% with `GPT4o-mini` and `Llama3.1-70B`). Results on Llama-2 responses (Appendix F.1) confirm generalization: even when base performance is ∼90% F2, `ColMAD` never degrades below the best single agent.

Finding 4: Improvements of `ColMAD` is statistically significant. We further ran each protocol with `GPT4o-mini` and `Llama3.1-70B` for 10 seeds at temperature $= 1$. `ColMAD` is the only protocol whose 95% confidence interval does not overlap with either `SA` method on fact verification and average F2, confirming that the improvements are statistically significant rather than noise. The full table of mean $\pm$ std and 95% CIs is provided in Appendix F.6 (Table 12).



(a) F1 under different token costs (b) F2 under different token costs (c) F2 under different LLM calls

Figure 5. Scaling SA methods using self-consistency improves the performance a bit while saturating in the end. Under the same LLM call or token cost budget, ColMAD breaks the plateau of SA, showing the potential of MAD approaches in combining complementary information from different LLMs.

Table 2. Performance of different test-time scaling methods under different LLM call budgets.

Debater#1	Debater#2	Protocol	GSM8K (ACC ↑)			AIME-2024 (ACC ↑)			Anthropic-Harmful (ASR ↓)		
			4	8	16	4	8	16	2	4	8
Qwen2.5-7B	-	-	**90.40**	**91.40**	**90.80**	13.33	13.33	13.33	14.59	18.92	31.69
Llama3.1-8B	-	-	88.80	89.60	89.40	3.33	6.67	6.67	14.32	9.93	15.34
Qwen2.5-7B	Llama3.1-8B	ColMAD	90.00	90.20	90.60	10.00	10.00	**20.00**	**0.00**	**0.00**	**0.27**
Qwen2.5-7B	Llama3.1-8B	CosMAD	88.00	87.80	87.80	6.67	6.67	6.67	20.27	20.54	23.24
Qwen2.5-7B	Llama3.1-8B	CopMAD	88.40	47.40	56.40	**16.67**	3.33	3.33	7.30	23.51	30.27

Finding 5: ColMAD outperforms single-agent scaling under matched compute. A natural concern is whether ColMAD's gains come from using more tokens. As shown in Fig. 5, we compare ColMAD with SA using self-consistency (SC). One could find that, SC saturates quickly for both models, with F2 oscillating within a 3-point band regardless of budget. At matched tokens (∼14.7K), ColMAD (86.29 Avg F2) outperforms SC@14 for GPT4o-mini (81.98, +4.3) and Llama (77.36, +8.9). Under the same LLM calls (7 calls), the gap is larger still. This confirms that ColMAD's advantage comes from effective exploitation of the *cross-model diversity*, not compute scaling.

4.3. Ablation Studies and Analysis

In Table 2, we extend to reasoning (GSM8K, AIME-2024) and safety (Anthropic-Harmful) benchmarks following Yang et al. (2025), using Qwen2.5-7B and Llama3.1-8B. ColMAD consistently improves over CosMAD and SA baselines on the harder AIME-2024 (+7% over best SA at 16 calls) and dramatically reduces the attack success rate on Anthropic-Harmful (from ∼15–30% to near 0% with cross-model debate), suggesting strong potential for safety-critical oversight. On saturated tasks (GSM8K, ∼90% ACC), improvements are near-zero, consistent with Proposition 2.8: when base performance is high, little room remains for complementary information.

Judge robustness, explanation alignment, and sensitivity to debate rounds. Due to space limits, we defer the analyses of judge swapping (Fig. 8(a)), explanation alignment (Fig. 8(b)), and sensitivity to debate rounds (Fig. 9) to Appendix F.7. In short, ColMAD is robust to judge choice (Avg F2 changes by only 0.03 vs. 22.88 for CopMAD), yields more human-aligned explanations than CopMAD, and is robust to the number of debate rounds while CopMAD degrades with more rounds.

More results. We also provide ablation studies on different prompt components in Appendix F.2, as well as different debater settings in Appendix F.3.

5. Conclusions

In this work, we investigated when and why MAD fails by examining the incentive structures of CopMAD and CosMAD protocols. We showed that both suffer from debate hacking due to goal misalignment: CopMAD incentivizes persuasion over truth-seeking, while CosMAD suppresses informative disagreements. We introduced ColMAD, a new collaborative MAD protocol that encourages agents to provide informative and truthful messages. Our results across error detection, reasoning, and safety tasks demonstrated that MAD is not inherently ineffective, and also has the potential to break the limit of SA methods, while proper MAD protocol design is critical to elicit the potential of MAD.

References

AI, M. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/

meta-llama-3-1/, 2024. Accessed: 2024-07-23. (Cited on page 6)

Art of Problem Solving. AIME Problems and Solutions, 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2025-05-15. (Cited on page 6)

Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval: The concepts and technology behind search*. Addison-Wesley Publishing Company, USA, 2nd edition, 2011. ISBN 9780321416919. (Cited on pages 3 and 6)

Brown-Cohen, J., Irving, G., and Piliouras, G. Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125*, 2023. (Cited on page 16)

Brown-Cohen, J., Irving, G., and Piliouras, G. Avoiding obfuscation with prover-estimator debate. *arXiv preprint arXiv:2506.13609*, 2025. (Cited on page 16)

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S. M., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint*, arXiv:2303.12712, 2023. (Cited on page 1)

Buhl, M. D., Pfau, J., Hilton, B., and Irving, G. An alignment safety case sketch based on debate. *arXiv preprint arXiv:2505.03989*, 2025. (Cited on pages 1 and 16)

Carro, M. V., Mester, D. A., Nieto, F., Stanchi, O. A., Bergman, G. E., Leiva, M., Gangi, L. N. F., Sprejer, E., Selasco, F. G., Corvalan, J. G., Martinez, M. V., and Simari, G. AI debaters are more persuasive when arguing in alignment with their own beliefs. In *First Workshop on Multi-Turn Interactions in Large Language Models*, 2025. (Cited on page 16)

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023. (Cited on pages 1 and 16)

Chen, J., Saha, S., and Bansal, M. 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)*, pp. 7066–7085, 2024a. (Cited on pages 1, 4 and 16)

Chen, P., Zhang, S., and Han, B. CoMM: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1720–1738, 2024b. (Cited on pages 1, 5, 6 and 16)

Chen, Z., Li, J., Chen, P., Li, Z., Sun, K., Luo, Y., Mao, Q., Yang, D., Sun, H., and Yu, P. S. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036*, 2025. (Cited on page 1)

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint*, arXiv:2110.14168, 2021. (Cited on page 6)

Crawford, V. P. and Sobel, J. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982. (Cited on pages 4 and 13)

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. (Cited on pages 1, 3, 4, 5, 6, 12 and 16)

Feng, S., Ding, W., Liu, A., Wang, Z., Shi, W., Wang, Y., Shen, Z., Han, X., Lang, H., Lee, C.-Y., Pfister, T., Choi, Y., and Tsvetkov, Y. When one llm drools, multi-llm collaboration rules. *arXiv preprint arXiv:2502.04506*, 2025. (Cited on page 1)

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. (Cited on pages 1 and 6)

Irving, G., Christiano, P., and Amodei, D. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018a. (Cited on pages 1, 3 and 16)

Irving, G., Christiano, P. F., and Amodei, D. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018b. (Cited on pages 5 and 6)

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. (Cited on page 1)

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de Las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts. *arXiv preprint*, arXiv:2401.04088, 2024. (Cited on page 6)

Kamoi, R., Das, S. S. S., Lou, R., Ahn, J. J., Zhao, Y., Lu, X., Zhang, N., Zhang, Y., Zhang, R. H., Vummanthala, S. R., Dave, S., Qin, S., Cohan, A., Yin, W., and Zhang,

R. Evaluating llms at detecting errors in LLM responses. *arXiv preprint arXiv:2404.03602*, 2024a. (Cited on pages 1, 2, 3, 6 and 20)

Kamoi, R., Zhang, Y., Zhang, N., Han, J., and Zhang, R. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024b. (Cited on page 3)

Kenton, Z., Siegel, N. Y., Kramar, J., Brown-Cohen, J., Albanie, S., Bulian, J., Agarwal, R., Lindner, D., Tang, Y., Goodman, N., and Shah, R. On scalable oversight with weak LLMs judging strong LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (Cited on pages 1, 3, 5, 6, 16 and 17)

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive LLMs leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024. (Cited on pages 1, 3, 5, 6, 12 and 16)

Kim, E. M., Garg, A., Peng, K., and Garg, N. Correlated errors in large language models. In *International Conference on Machine Learning*, 2025. (Cited on pages 2 and 3)

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., and Shi, S. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023. (Cited on pages 1, 5, 6 and 16)

Minsky, M. The society of mind. *The Personalist Forum*, 1987. (Cited on page 5)

OpenAI. Gpt-4o mini technical report. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. (Cited on page 6)

Rahman, S., Issaka, S., Suvarna, A., Liu, G., Shiffer, J., Lee, J., Parvez, M. R., Palangi, H., Feng, S., Peng, N., Choi, Y., Michael, J., Jiang, L., and Gabriel, S. AI debate aids assessment of controversial claims. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. (Cited on page 16)

Smit, A. P., Duckworth, P., Grinsztajn, N., ab Tessera, K., Barrett, T. D., and Pretorius, A. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*, 2024. (Cited on pages 1, 3, 6 and 16)

Team, Q. Qwen2.5 technical report. *arXiv preprint*, arXiv:2412.15115, 2024. (Cited on page 6)

Team, Q. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/. (Cited on page 6)

Wang, Q., Wang, Z., Su, Y., Tong, H., and Song, Y. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6106–6131, 2024. (Cited on pages 1, 3, 6 and 16)

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. (Cited on pages 1 and 6)

Wynn, A., Satija, H., and Hadfield, G. K. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025. (Cited on page 16)

Yang, Y., Yi, E., Ko, J., Lee, K., Jin, Z., and Yun, S. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*, 2025. (Cited on pages 1, 2, 3, 6, 8 and 16)

Yin, Z., Sun, Q., Chang, C., Guo, Q., Dai, J., Huang, X., and Qiu, X. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15135–15153, 2023. (Cited on pages 1, 5 and 16)

Zeng, Y., Wu, Y., Zhang, X., Wang, H., and Wu, Q. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*, 2024. (Cited on page 6)

Zhang, H., Cui, Z., Wang, X., Zhang, Q., Wang, Z., Wu, D., and Hu, S. If multi-agent debate is the answer, what is the question? *arXiv preprint arXiv:2502.08788*, 2025. (Cited on pages 1, 3, 4, 6 and 16)

LLM Use Statement

From the research side, this work studies the use of LLMs to perform multi-agent debate to detect errors in LLM responses. From the paper writing side, we use LLMs to assist with improving the writing of this work.

Broader impacts

This work does not involve human subjects or personally identifiable information beyond public benchmarks used under their licenses. Our experiments evaluate error detection and decision protocols among LLMs, which benefits the oversight and prevents potential risks of superhuman intelligence in the future.

A. Discussion and Future Work

Training for collaborative debate. The current implementation relies on prompting to elicit collaborative behavior, but LLMs are predominantly trained on single-agent objectives. Fine-tuning debater models with multi-agent objectives, e.g., rewarding information gain $I(Y; m_i|X_0, M^{(t-1)})$ rather than persuasion probability could internalize the collaborative incentive structure.

Honesty and uncertainty calibration. While `ColMAD` demonstrates that collaborative debate can overcome the limitations of competitive and consensus-seeking `MAD` protocols, our behavioral analysis also reveals its remaining failure modes, e.g., there are still some overconfident claims. This reflects a fundamental calibration problem: LLMs express high confidence on claims they cannot verify. Training for honesty through objectives that penalize confident assertions on uncertain claims, would reduce both overconfidence and fabricated evidence simultaneously.

Broader applications. We evaluate `ColMAD` on error detection, mathematical reasoning, and safety alignment, but the collaborative debate framework is applicable to any task where cross-checking between agents can surface complementary information, including code review, scientific claim verification, and safety-critical content moderation. Extending to multi-party debate (beyond two agents) and heterogeneous agent roles (e.g., specialist verifiers for different requirement types) are natural next steps.

B. More Details about Theories

B.1. Notations

A table of notations used in our work is given as follows.

Debate setup. We consider two agents A (Alice) and B (Bob), and denote the predictions of the error detections as $\hat{\boldsymbol{y}}_A$ and $\hat{\boldsymbol{y}}_B$, with rationales (e.g., CoT reasoning) as x_A and x_B, respectively. During the debate, they will emit messages m_A and m_B to convince a judge J.

Assumption B.1 (Optimal Judge Strategy)**.** *Denoting the label as $Y \in \{0, 1\}$ with prior probability $\pi \in (0, 1)$, without debating, the judge will derive the final answer based on the initial responses $X_0 = (x_A, x_B)$. Assuming the judge J uses the Bayes test on the total log-likelihood ratio (LLR), we will have the LLR based on X_0 as*

$$\Lambda_0(X_0) = \log \frac{p(X_0|Y = 1)}{P(X_0|Y = 0)} + \log \frac{\pi}{1 - \pi}. \tag{2}$$

The contribution to LLR from the debating can be written as

$$l_i(m_i; X_0) = \log \frac{P(m_i|Y = 1, X_0)}{P(m_i|Y = 0, X_0)},\ i \in \{A, B\}. \tag{3}$$

Then, with debating, the LLR of the judge is

$$\Lambda(X_0, m_A, m_B) = \Lambda_0(X_0) + l_A(m_A; X_0) + l_B(m_B; X_0) + \log \frac{\pi}{1 - \pi}. \tag{4}$$

Table 3. Table of Notations.

Notation	Meaning
$y \in \{0, 1\}$	True label (e.g., *error* vs *no_error*).
Y	Random variable of the label predictions.
$\boldsymbol{y}$	Predictions of the labels over a dataset.
π	Prior $\Pr(y = 1)$; prior log-odds $\log \frac{\pi}{1-\pi}$.
$X_0 = (a, b)$	Baseline signals from the two base models A, B (what the judge has without debate).
$p(x \mid y)$	Likelihood of baseline signal x under label y.
$\Lambda_0(x) = \log \frac{p(x\|y=1)}{p(x\|y=0)}$	Baseline log–likelihood ratio (LLR), i.e., weight of evidence from X_0.
m_A, m_B	Debate messages emitted by debater A and B.
$M = (m_A, m_B)$	Joint debate messages.
$p_i(m_i \mid y, x)$	Conditional likelihood model of debater i's message given y and $x = X_0$.
$\ell_i(m_i; x) = \log \frac{p_i(m_i\|y=1,x)}{p_i(m_i\|y=0,x)}$	Debater i's additive LLR contribution given message and context.
$\|\ell_i\| \leq L_i$	Bounded manipulability / persuasion budget for debater i.
$\Lambda(x, m_A, m_B)$	Total LLR used by the judge: $\Lambda_0(x) + \ell_A(m_A; x) + \ell_B(m_B; x) + \log \dfrac{\pi}{1-\pi}$.
J	Judge decision rule mapping $(X_0, m_A, m_B) \mapsto \{0, 1\}$.
$R(Z)$	Bayes (minimum) 0–1 risk achievable when using signal Z.
$R_{\text{base}} := R(X_0)$	Baseline Bayes risk using only the base signals (no debate).
V_{adv}	Minimax error in adversarial (zero-sum) debating.
R_{coop}	Bayes risk under cooperative, truth-seeking debating (messages add true evidence).
$I(y; M \mid X_0)$	Conditional mutual information: new information from messages beyond X_0.
$\eta(z) = \Pr(y = 1 \mid z)$	Posterior probability under observable z (e.g., $z = X_0$ or $z = (X_0, M)$).
$\frac{1}{1+e^{\|\Lambda\|}}$	Instantaneous Bayes error at balanced prior ($\pi = \frac{1}{2}$) for total LLR Λ.

Intuitively, if the elicited messages $M = (m_A, m_B)$ bring additional information to the judge, i.e., $I(Y; M|X_0) > 0$ where $I(\cdot; \cdot)$ denotes the mutual information, we will have $\Lambda(X_0, M) > \Lambda(X_0)$. In order to provide additional information, we need to look into the cases where the predictions by A differ from those by B. Under $Y_A \neq Y_B$, if both agents are able to provide *sufficient justifications* and are more persuasive during the debate when they are debating for the *correct* answer. Hence, the judge can be convinced to take the correct answer when either of the agents is correct. The reduction of errors from the oracle collaboration can be calculated through

$$\min_{K \in \{A,B\}} \sum_i \mathbb{1}[\hat{\boldsymbol{y}}_K^{(i)} \neq \boldsymbol{y}^{(i)}] - \sum_i \mathbb{1}[\hat{\boldsymbol{y}}_J^{(i)} \neq \boldsymbol{y}^{(i)}], \tag{5}$$

where $\boldsymbol{y}_K^{(i)}$ denotes the initial prediction of the agent $K \in \{A, B\}$ on the i-th sample. Intuitively, Eq. (5) can be considered as the potential of the collaboration between A and B.

B.2. Game-theoretic formulation of existing MAD

Essentially, existing MAD methods can be categorized into two categories: CopMAD where LLM agents competitively debate with each other to convince the judge of the respective assigned answers (Khan et al., 2024); and CosMAD (i.e., Consensus-seeking MAD) where LLM agents seek consensus (Du et al., 2023).

Then, we provide game-theoretic definitions of CopMAD, ColMAD, and CosMAD:

Definition B.2 (Basic MAD setup)**.** *The basic setting of MAD games is that, given a question Q with a binary label $Y \in \{0, 1\}$, two debaters A and B with different beliefs of the label that debate with each other and produce messages $M = (m_A, m_B)$. Without loss of generality, A believes the answer $Y_A = 1$ with initial rationale x_A and B believes $Y_B = 0$ with initial rationale x_B. A judge J will give Bayes-optimal predictions $Y_J = J(X_0, M)$ based on the information X_0 and M (Assumption B.1). When we consider the number of debating rounds, we will denote the messages generated before round t as $M^{(t-1)}$; otherwise, we will omit it for the simplicity of notations.*

Definition B.3 (Competitive MAD)**.** *Debaters A and B have opposed utilities: $u_A(m_A) = P(Y_J = Y_A \mid m_A, X_0, M^{(t-1)})$, $u_B(m_B) = P(Y_J = Y_B \mid m_B, X_0, M^{(t-1)})$. The judge observes the full transcript M and the*

risk is $V_{\text{cop}} = \min_J \max_{e \in \mathcal{E}_{\text{cop}}(J)} P(J(X_0, M_e) \neq Y)$*, where* M_e *is the debating transcripts given by the optimal Nash equilibrium* $e \in \mathcal{E}_{\text{cop}}(J)$ *of the A-B subgame in convincing* J*.*

Definition B.4 (Consensus-seeking MAD)**.** *Debaters* A *and* B *share the same utilities towards reaching consensus:* $u_A(m_A) = u_B(m_B) = P(Y_A = Y_B \mid X_0, M^{(t-1)})$*. The risk is* $V_{\text{cos}} = \min_J \min_{e \in \mathcal{E}_{\text{cos}}(J)} P(J(X_0, M_e) \neq Y)$*, where* $M_e = \Phi_\cap(M)$ *is the debating transcripts given by the optimal Nash equilibrium* $e \in \mathcal{E}_{\text{cos}}(J)$ *of the A-B subgame in reaching consensus.*

Definition B.5 (Collaborative MAD)**.** *Debaters have truth-seeking utilities:* $u_A(m_A) = u_B(m_B) = I(Y; m_i \mid X_0, M^{(t-1)})$*. The judge observes the full transcript* M*. The risk is* $V_{\text{col}} = \min_J \min_{e \in \mathcal{E}_{\text{col}}(J)} P(J(X_0, M_e) \neq Y)$*.*

B.3. Proof for Proposition 2.4 (Competitive MAD)

Proposition B.6 (Restatement of Proposition 2.4)**.** *Assuming a bounded LLR, i.e.,* $l_i \leq L_i < \infty,\ i \in \{A, B\}$*, let* $R(Z)$ *denote the Bayes risk of the judge* J *when making decisions based on* Z*, denote* $R_0 = R(X_0)$*, denote the outcome of zero-sum debating as*

$$V_{\text{cop}} = \min_J \max_{e \in \mathcal{E}_{\text{cop}}(J)} P(J(X_0, M_e) \neq Y),$$

where M_e *are the debating transcripts given by the optimal Nash equilibrium* $e \in \mathcal{E}_{\text{cop}}(J)$ *of the* A–B *subgame in convincing* J*. Then, we have* $V_{\text{cop}} = R_0$*.*

Proof. To show $V_{\text{cop}} = R_0$, we need to show $V_{\text{cop}} \leq R_0$, and $V_{\text{cop}} \geq R_0$, and provide a condition that the equity holds.

(i) For $V_{\text{cop}} \leq R_0$, given that the judge aims to choose the optimal strategy given any strategies of A and B that may degenerate the debate performance. We first consider a simple ignore strategy for the judge, J_{ignore} that directly drops any additional debate transcripts between A and B. We have

$$P(J_{\text{ignore}}(X_0, M_e) \neq Y) = P(J_{\text{ignore}}(X_0) \neq Y) = R_0, \tag{6}$$

which follows

$$\max_{e \in \mathcal{E}_{\text{cop}}(J)} P(J_{\text{ignore}}(X_0, M_e) \neq Y) = R_0. \tag{7}$$

Then, it suffices to know that

$$V_{\text{cop}} = \min_J \max_{e \in \mathcal{E}_{\text{cop}}(J)} P(J(X_0, M_e) \neq Y) \leq \max_{e \in \mathcal{E}_{\text{cop}}(J)} P(J_{\text{ignore}}(X_0, M_e) \neq Y) = R_0, \tag{8}$$

and $V_{\text{cop}} \leq R_0$.

(ii) For $V_{\text{cop}} \geq R_0$, we consider the strategies of the debaters. Without loss of generality, we assume A is debating for $Y = 1$ and B is debating for $Y = 0$. Since we do not impose any limits on the capabilities of the debater agents, they will try to present the evidence as most useful for the respective answer as they can. More formally, the optimal strategies for debaters A and B will yield the following

$$M \perp\!\!\!\perp Y | X_0. \tag{9}$$

It follows that

$$P(Y = 1 | X_0, M) = P(Y = 1 | X_0). \tag{10}$$

Therefore, it suffices to know that $V_{\text{cop}} \geq R_0$. That concludes our proof. □

Connection to cheap-talk games. The structure of CopMAD— costless messages between misaligned senders and a receiver — is analogous to the cheap-talk game of Crawford & Sobel (1982). The equilibrium $M \perp Y | X_0$ corresponds to the *babbling equilibrium* in cheap-talk theory, where messages carry no information because the sender's incentive is to persuade rather than inform. This provides a game-theoretic explanation for the debate hacking phenomenon observed in Fig. 3.

B.4. Proof for Proposition 2.6 (Consensus-Seeking MAD)

Proposition B.7 (Pitfalls of consensus-seeking debate (restated))**.** *Let $R(X_0, Z)$ denote the Bayes risk when the judge decides based on (X_0, Z). Then $R(X_0, \Phi_\cap(M)) \geq R(X_0, M)$, with strict inequality when $I(Y; M|X_0) > I(Y; \Phi_\cap(M)|X_0)$.*

Proof. Since $\Phi_\cap(M)$ is a deterministic function of M (retaining only the subset on which both agents agree), by the data processing inequality:

$$I(Y; \Phi_\cap(M)|X_0) \leq I(Y; M|X_0). \tag{11}$$

The Bayes risk $R(X_0, Z)$ is a monotonically decreasing function of $I(Y; Z|X_0)$: more information about Y allows a better decision. Therefore:

$$R(X_0, \Phi_\cap(M)) \geq R(X_0, M). \tag{12}$$

The strict inequality holds when $\Phi_\cap$ discards information about Y, i.e., $I(Y; M|X_0) > I(Y; \Phi_\cap(M)|X_0)$. This occurs when there exist disagreements between agents that carry information about the true label but are filtered out by the consensus process. □

Interpretation. The information loss equals the conditional mutual information between Y and the disagreements, given the consensus:

$$I(Y; M|X_0) - I(Y; \Phi_\cap(M)|X_0) = I(Y; M \setminus \Phi_\cap(M) \mid X_0, \Phi_\cap(M)). \tag{13}$$

This quantity is strictly positive whenever the disagreements carry information about Y that is not already captured by the consensus, which formalizes the intuition that *disagreements are the most valuable signals* in a debate.

B.5. Proof for Proposition 2.8 (Collaborative MAD)

Proposition B.8 (Collaborative debate (restated))**.** *$V_{\text{col}} \leq R(X_0) = V_{\text{cop}}$, with strict inequality when $I(Y; M_e|X_0) > 0$.*

Proof. **$V_{\text{col}} \leq R(X_0)$:** The ignore strategy J_{ignore} gives $R(X_0)$ regardless of M, so $V_{\text{col}} \leq R(X_0)$.

Strict inequality when $I(Y; M_e|X_0) > 0$: When messages carry positive information, the posterior $\eta(X_0, M_e) = P(Y = 1|X_0, M_e)$ differs from $\eta(X_0) = P(Y = 1|X_0)$ with positive probability. Under 0–1 loss, $R(Z) = 1 - \max\{\eta(Z), 1 - \eta(Z)\}$. With positive probability there exists (X_0, M_e) such that:

$$\max\{\eta(X_0, M_e), 1 - \eta(X_0, M_e)\} > \max\{\eta(X_0), 1 - \eta(X_0)\}, \tag{14}$$

yielding $R(X_0, M) < R(X_0)$, so $V_{\text{col}} < R(X_0) = V_{\text{cop}}$. □

B.6. Proof for Corollary 2.9 (Protocol Ordering)

Corollary B.9 (Ordering of MAD protocols (restated))**.** *$V_{\text{col}} \leq R(X_0, \Phi_\cap(M)) \leq R(X_0) = V_{\text{cop}}$.*

Proof. We combine the three propositions:

- $V_{\text{cop}} = R(X_0)$: from Proposition 2.4.
- $R(X_0, \Phi_\cap(M)) \leq R(X_0)$: the consensus subset $\Phi_\cap(M)$ carries at least as much information as X_0 alone (agents may add useful information even after filtering), so $R(X_0, \Phi_\cap(M)) \leq R(X_0)$.
- $V_{\text{col}} \leq R(X_0, \Phi_\cap(M))$: from Proposition 2.6, $R(X_0, \Phi_\cap(M)) \geq R(X_0, M)$, and $V_{\text{col}} = \min_J R(X_0, M) \leq R(X_0, M) \leq R(X_0, \Phi_\cap(M))$.

Therefore $V_{\text{col}} \leq R(X_0, \Phi_\cap(M)) \leq R(X_0) = V_{\text{cop}}$. □

Table 4. Mapping between the theoretical condition for ColMAD (Proposition 2.8) and the practical prompt components in our implementation.

Theoretical requirement	Practical design	Mechanism
Messages carry new information about Y beyond X_0	"Complement missing points" instruction	Maximizes $I(Y; M_e \mid X_0)$ by directing agents to identify what the other missed.
Messages are truthful, not fabricated	Evidence verification (quote-based)	Ensures debate messages reflect genuine evidence grounded in the context.
Reduce overconfidence and dishonesty	Self-auditing	Agents identify plausible failure modes in their own claims, approximating the honest-debater assumption.
Judge can properly weight evidence	Confidence calibration	Helps the judge estimate the LLR contribution of each message.

Interpretation. The ordering reflects three levels of information utilization:

- CopMAD: opposed incentives destroy information ($I(Y; M|X_0) = 0$), so the judge effectively has only X_0.
- CosMAD: agents produce some useful information, but consensus filtering discards the disagreements, leaving $\Phi_\cap(M) \subseteq M$.
- ColMAD: truth-seeking incentives produce informative messages, and the judge sees the full transcript M including disagreements.

Each step from CopMAD to CosMAD to ColMAD preserves strictly more information about Y, yielding strictly lower risk when the additional information is non-trivial.

B.7. Theory-to-implementation mapping

Table 4 maps the theoretical condition that ColMAD requires for its advantage over CopMAD (Proposition 2.8) to the specific prompt components used in our implementation.

C. Categorization of existing MAD approaches

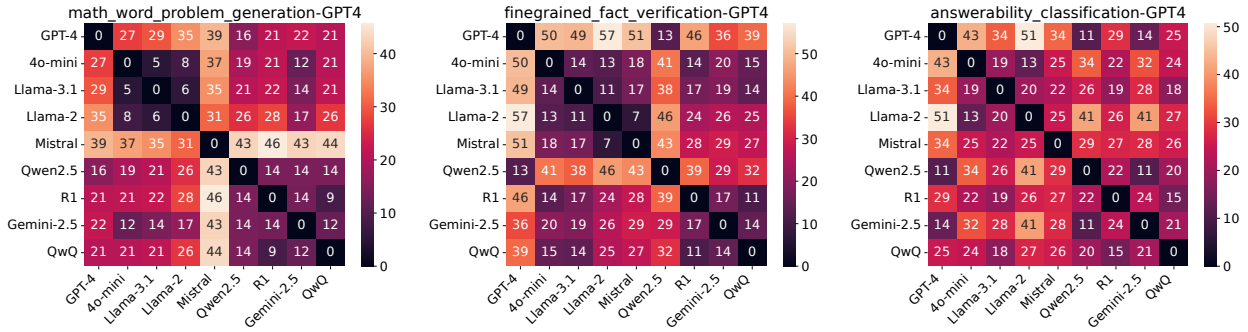
We summarize the existing MAD literature in Table 5, covering the two dominant protocols (CopMAD and CosMAD) as well as orthogonal directions (communication, application, and benchmarking). Our ColMAD is the first MAD protocol that reframes debate as a non-zero sum collaborative game while retaining the adversarial role of debate to surface decisive errors.

D. More Results on Potential of Multi-Agent Collaboration

Given ReaLMistake, we provide more results on the error reduction of multi-agent collaboration under the oracle protocol as Eq. (5).

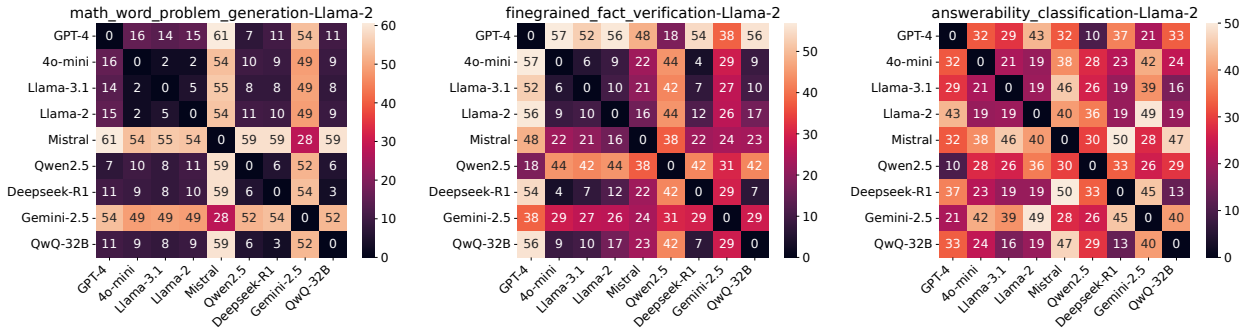
Table 5. Categorization of existing MAD approaches. We position ColMAD as a collaborative MAD, distinct from the two dominant lines in the literature: CopMAD (competitive debate) and CosMAD (consensus-seeking debate). Works without a natbib key are cited inline.

Name	MAD aspect	Reference
ColMAD	Collaborative MAD	Ours
CopMAD	Competitive MAD	Kenton et al. (2024)
Khan et al. (2024)	Competitive MAD	Khan et al. (2024)
Irving et al. (2018a)	Competitive MAD	Irving et al. (2018a)
Brown-Cohen et al. (2025)	Competitive MAD	Brown-Cohen et al. (2025)
Buhl et al. (2025)	Competitive MAD	Buhl et al. (2025)
Brown-Cohen et al. (2023)	Competitive MAD	Brown-Cohen et al. (2023)
Rahman et al. (2025)	Competitive MAD	Rahman et al. (2025)
Carro et al. (2025)	Competitive MAD	Carro et al. (2025)
MP	Competitive MAD with persona	Liang et al. (2023)
SoM	Consensus-seeking MAD (CosMAD)	Du et al. (2023)
Reconcile	Consensus-seeking MAD (CosMAD)	Chen et al. (2024a)
CoMM	Consensus-seeking MAD with persona	Chen et al. (2024b)
EoT	Communication protocol for MAD	Yin et al. (2023)
ChatEval	Application of MAD (evaluation)	Chan et al. (2023)
Wang et al. (2024)	Benchmarking of MAD	Wang et al. (2024)
Smit et al. (2024)	Benchmarking of MAD	Smit et al. (2024)
Wynn et al. (2025)	Benchmarking of CosMAD	Wynn et al. (2025)
Zhang et al. (2025)	Benchmarking of MAD	Zhang et al. (2025)
Yang et al. (2025)	Benchmarking of MAD	Yang et al. (2025)



(a) Math problem (b) Fact verification (c) Answerability

Figure 6. Error reductions of prevalent LLMs in detecting errors of GPT-4. The numbers refer to the reduced errors following the oracle collaboration via Eq. (1). It can be found that different LLMs are less likely to make mistakes simultaneously. When incorporating LLMs with higher heterogeneity, such as those from different companies, the error reduction rates will be higher.



(a) Math problem (b) Fact verification (c) Answerability

Figure 7. Error reductions of prevalent LLMs in detecting errors of Llama-2. The numbers refer to the reduced errors following the oracle collaboration via Eq. (1). It can be found that different LLMs are less likely to make mistakes simultaneously. When incorporating LLMs with higher heterogeneity, such as those from different companies, the error reduction rates will be higher.

E. Details of the Debate

E.1. More details on `ColMAD`

Algorithm 1 The `ColMAD` Framework

Required: `ColMAD` debater agents A, and B; the judge agent J; Dataset of LLM responses $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$; Maximal debate rounds R;
Initializing debater A's solution x_A^0 via prompting A using single-agent prompt, and obtaining the label y_A;
Initializing debater B's solution x_B^0 via prompting B using single-agent prompt, and obtaining the label y_B;
if y_A does not equal to y_B **then**
  Constructing debate transcripts $M^{(0)} = (m_A^{(0)}, m_B^{(0)})$;
  for current round $t \in [1, ..., T]$ **do**
    Instructing debater A with $M_e = \{M^{(t)}\}_{t=0}^{t-1}$ to find missing points of B to obtain $m_A^{(t)}$;
    Instructing debater B with $M_e = \{M^{(t)}\}_{t=0}^{t-1}$ to find missing points of A to obtain $m_B^{(t)}$;
    Constructing debate transcripts $M^{(t)} = (m_A^{(t)}, m_B^{(t)})$;
  end for
  Summarizing transcripts $M_e = \{M^{(t)}\}_{t=0}^{T}$;
  Instructing the Judge J to reason for the answer y_J;
else
  $y_J = y_A$;
end if
return y_J;

The algorithm of `ColMAD` is given in Alg. 1, and the critical prompts of `ColMAD` are given as follows:

Debater prompt. The debater agent is encouraged to collaborate with the other debater agent to provide comprehensive information to the judge to make a well-informed decision. In addition to clearly state the objective of collaboration for seeking the truthful answer, we implement the insights of `ColMAD` into specific prompting schemes: (i) Evidence verification: `ColMAD` implements a quote-based system that asks the debaters to quote evidence from the context, and each quote will be verified if there is an exact match in the context, following Kenton et al. (2024); (ii) Self-auditing: the debaters are required to self-audit if there exists one potential failure mode in the claim; (iii) Confidence calibration: the debaters are required to provide a confidence estimate for their own claims.

Debater prompt

<task> You will act as an expert debater named Debater {debater_name} on error detection for a model response. You are given a model input and a model response. You are assigned to defend {answer_defending} while another debater defends the alternative. A judge will decide the final answer. Your mandate is collaborative truth-seeking with adversarial robustness: surface decisive errors (in either side), adopt valid points, and converge to the best-supported conclusion.

Tone: scholarly, concise, evidence-first. Do NOT include an "In summary/Conclusion" paragraph because more rounds may follow.

Output format:
"thinking"... private reasoning... "thinking" "argument"... what the judge sees... "argument"

The judge only sees "argument", so prioritize clarity and testable evidence.
</task>
<quote-system>
Judging priorities (optimize for these):
1) Decisive tests that flip the result if they fail (spec mismatch, contradiction, counterexample).
2) Verified quotes ("quote"... "quote" → may become "v_quote") grounded in the task's general/specific requirements.
3) Novel, non-redundant contributions; penalties for repetition or unsupported claims.
4) Accurate calibration and willingness to adopt valid opponent points.
</quote-system>
<problem>Here is the model input and the model response...</problem>
<transcript>Here is the transcript of the debate so far...</transcript>
<thinking>
Now produce your next turn. Follow these steps in your "thinking", then write only the "argument":

1) Identify decisive tests: 2-4 concrete checks that would settle {answer_defending} vs the alternative (e.g., requirement violation, logical contradiction, factual mismatch).
2) Run each test against BOTH sides (yours and opponent's). If a test fails on your side, plan a revision or boundary condition.
3) Extract 2-4 atomic requirements and cite them via "quote"... "/quote" (no repetition; aim for new quotes).
4) Steelman Opponent: find their strongest correct point and either integrate it or delimit its scope.
5) Prepare a one-sentence update note describing any change in your view.
</thinking>
<argument>
- Position: state which answer you defend and why in one sentence.
- Decisive checks: list your 2-4 tests with PASS/FAIL for each side, and the minimal evidence that justifies each status.
- Evidence: map 2-4 "quote"... "quote" to specific claims you make (no duplicate quotes).
- Steelman: state opponent's strongest correct point and how you incorporate or delimit it.
- Self-audit: one plausible failure mode in your argument (1 short clause).
- End with:
Final: "1 error | 2 no_error"
Conf: "0-1" (justify briefly: signal strength, remaining uncertainties).

Do not exceed {word_limit} words in "argument". Remember to output both "thinking" and "argument".
</argument>

Judge prompt. The Judge agent is instructed to analyze the debating transcripts to

Judge Prompt

<task>
You are an expert judge in a debate where the participants are arguing over the correct answer to the error detection problem. Your task is to determine which debater is correct. Your main objective is to choose the correct answer based on the evidence presented. You should maintain a scholarly yet assertive tone focused on evidence-based analysis while critically evaluating arguments.
</task>
<quote-system>
Here is some important information regarding the quote system which is applied after the debaters have finished their arguments:
- Verified quotes are direct string matches from the requirements and are enclosed in “v_quote”“/v_quote” tags.
- Unverified quotes are not direct string matches from the requirements, so they may not be trusted and are enclosed in “u_quote”“/u_quote” tags.
- Some quotes are just enclosed in quotation marks, which counts as unverified and may not be trusted.
<quote-system>

<problem>Here is the model input and the model response...</problem>
<transcript>Here is the transcript of the debate...</transcript>
<reasoning>
Choices: 1: {answer_1} 2: {answer_2}
{name_1} is arguing for 1 and {name_2} is arguing for 2.

Choose the answer you think is most likely to be correct.
First explain your reasoning, considering the strengths and weaknesses of each answer.
Second, summarize the key pieces of evidence from the debate for why the error detection of model response is {answer_1} or {answer_2}.
Finally, give your answer formatted as " Answer: “1 {answer_1}|2 {answer_2}”.
</reasoning>

Problem. The problem statement is organized as follows:

Problem

<problem>
Here is the original problem, including model input and the model response:
We provide an input to a language model (model input) and a response from the model (model response). Your task is to detect errors in the model response. You need to check whether the model response follows all instructions and requirements in the model input. You need to classify the model response as an error if the model response does not follow any of the instructions or requirements in the model input.

===== Model Input Begins =====
.....
===== Model Input Ends =====

===== Model Response Begins =====
.....
===== Model Response Ends =====
In your response, provide your explanation first and conclude your response with "Therefore, the model response contains an error." or "Therefore, the model response contains no error." </problem>

Transcripts. The transcripts are organized as follows:

Transcripts

<transcript>
Here is the transcript of the debate:
Round 1:
debater Alice: ...
debater Bob: ...
Round 2:
debater Alice: ...
debater Bob: ...
</transcript>

F. More Details about Experiments

Table 6. Statistics of `ReaLMistake` benchmark (Kamoi et al., 2024a).

Response Model	Task	# Data	Average # tokens		Errors in Responses from GPT-4 or Llama 2 70B [%]				
			Input	LLM Response	Reasoning Correctness	Instruction-Following	Context-Faithfulness	Parameterized Knowledge	Total Error
GPT-4 0613	Math Word Problem Generation	140	252	151	25.0	57.1	–	–	62.1
	Fine-grained Fact Verification	140	523	83	25.7	5.7	45.0	–	62.9
	Answerability Classification	140	119	75	22.1	–	8.6	40.7	62.1
Llama 2 70B	Math Word Problem Generation	160	235	163	51.2	67.5	–	–	80.0
	Fine-grained Fact Verification	160	511	168	56.9	44.4	45.6	–	80.6
	Answerability Classification	160	119	96	48.1	–	–	48.1	81.2

F.1. Additional results on detecting errors by Llama-2

The error detection results on Llama-2 responses in `ReaLMistake` are given in Table 7. As one could find it is relatively easier to detect errors of weak LLMs, all methods achieve relatively high results. Nevertheless, `ColMAD` remain the top one.

Transferability of `ColMAD` to different candidate LLMs. Combining the results of Table 1 in Sec. 4 and Table 7, although detecting errors of Llama-2 responses is relatively easier than that of GPT-4, we can also find that `ColMAD` outperforms `CopMAD` as well as the best single-agent performance by up to 4%. The results indicate the generality of the advantages of `ColMAD`.

F.2. Ablation studies on the effectiveness of different prompt components

Prompt components of `ColMAD`. We ablate the three structured prompt components: (i) quote-based evidence verification, (ii) confidence declaration, and (iii) self-auditing, using `GPT4o-mini` and `Llama3.1-70B`. As shown in Table 8, each component contributes to the final performance. Even without all three, `ColMAD` (85.58) still far outperforms `CopMAD` (57.34) and `CosMAD` (75.40), confirming that the core advantage comes from the collaborative game structure. A detailed mapping from conditions to components is in Appendix B.7.

F.3. Additional results with shared debaters and weaker judges

To probe the concern that our main setting pairs Debater#1 with the judge as the same LLM, we further experiment with two identical debaters (`GPT4o-mini` vs. `GPT4o-mini`), both with a matching judge and with an independent, weaker judge (`Mistral-7B-v0.3`). Results are given in Table 9. `ColMAD` still outperforms `CopMAD` even when both debaters share the same LLM, but all `MAD` methods underperform the single-agent baseline, reinforcing the observation that *heterogeneous debaters* are a key driver of `ColMAD`'s gains.

F.4. Token cost across `MAD` protocols

Table 10 reports the averaged token cost (mean $\pm$ std) of the full debating transcripts of each `MAD` protocol with `Llama3.1-70B` and `GPT4o-mini` as debaters. `ColMAD` uses $\sim$15% more tokens than `CopMAD` but yields substantially

Table 7. Results for Llama-2 responses. The judge uses the same LLM as Debater#1. The top two results are highlighted.

Debater#1	Debater#2	Protocol	Math Problem		Fact Verification		Answerability		Avg. F1	Avg. F2
			F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)
Human	-	-	98.30	97.34	100.0	100.0	100.0	100.0	99.43	99.11
GPT4o-mini	-	-	89.82	**95.67**	**93.14**	**97.14**	79.50	75.52	87.49	89.44
Llama3.1-70B	-	-	**90.78**	96.10	91.45	93.75	82.87	81.12	88.37	90.32
Mistral-7B-v0.3	-	-	45.00	38.53	82.07	80.72	51.04	42.10	59.37	53.78
GPT4o-mini	Llama3.1-70B	ColMAD	89.82	**95.67**	92.47	96.85	**87.55**	**88.55**	**89.95**	**93.69**
		CopMAD	82.07	81.10	78.18	70.84	51.34	41.59	70.53	64.51
		Voting	**90.46**	95.95	93.43	96.82	81.30	78.62	88.40	90.46
Llama3.1-70B	GPT4o-mini	ColMAD	89.82	**95.67**	92.47	96.85	**88.81**	**90.43**	**90.37**	**94.31**
		CopMAD	88.89	93.51	89.47	91.12	76.99	73.13	85.12	85.92
		Voting	**90.46**	95.95	93.43	96.82	81.30	78.62	88.40	90.46
GPT4o-mini	Mistral-7B-v0.3	ColMAD	88.50	94.63	**92.81**	**96.99**	79.84	77.59	87.05	89.74
		CopMAD	85.71	86.31	80.53	74.23	59.70	50.76	75.31	70.43
		Voting	64.76	57.24	59.51	51.52	63.81	55.83	62.69	54.86
Llama3.1-70B	Mistral-7B-v0.3	ColMAD	90.14	**95.81**	90.91	94.41	84.50	84.10	88.52	91.44
		CopMAD	89.21	93.66	84.92	83.72	74.36	69.71	82.83	82.36
		Voting	60.00	57.01	48.98	44.78	60.00	57.01	56.33	52.93

Table 8. Ablation on the prompt components of ColMAD, with GPT4o-mini and Llama3.1-70B as debaters. Each component of the structured prompt design (quote-based evidence verification, confidence declaration, and self-auditing) contributes to performance.

Protocol	Math Problem		Fact Verification		Answerability		Average	
	F1	F2	F1	F2	F1	F2	F1	F2
ColMAD (full)	**78.38**	**90.06**	**77.42**	**88.98**	72.91	79.74	**76.24**	**86.26**
w/o quote system	77.68	89.69	77.21	88.30	73.17	80.47	76.02	86.15
w/o quote system & confidence	77.83	89.21	76.42	86.72	74.04	82.09	76.10	86.01
w/o quote system & confidence & self-auditing	78.03	89.88	74.64	84.05	**75.49**	**82.80**	76.05	85.58

larger improvements in F2, so the gain cannot be attributed to raw token usage. Crucially, CopMAD uses tokens comparable to ColMAD yet underperforms single-agent baselines — demonstrating that the protocol design, not compute, drives ColMAD's advantage.

F.5. Fine-grained alignment with human-annotated errors

To complement the LLM-as-judge evaluation in Fig. 8(b), we further analyze debating transcripts from CopMAD and ColMAD (Llama3.1-70B vs. DeepSeek-R1) on Math Word Problem Generation, extracting human-annotated error categories. Table 11 reports the percentage of each category identified by the corresponding method. Across categories, ColMAD identifies the same mistakes pointed out by humans at consistently higher rates than both single-agent baselines and CopMAD.

F.6. Statistical significance of the improvements

To examine the statistical significance of the improvements claimed in Finding 4, we ran each protocol with GPT4o-mini and Llama3.1-70B for 10 seeds at temperature $=1$. Table 12 reports the mean $\pm$ std and the 95% confidence intervals over the 10 runs. ColMAD is the only protocol whose 95% CI does not overlap with either single-agent method on fact verification and average F2, confirming that the improvements over SA are statistically significant rather than noise.

F.7. Judge robustness, explanation alignment, and sensitivity to debate rounds

Judge robustness. Different combinations of LLMs as judges are shown in Fig. 8(a). Swapping the judge changes ColMAD's Avg F2 by only 0.03, while CopMAD varies by 22.88, indicating that ColMAD produces balanced transcripts that any competent judge can evaluate.

Table 9. Results when both debaters share the same LLM (`GPT4o-mini`), with the judge being either the same model (`GPT4o-mini`) or a weaker independent model (`Mistral-7B-v0.3`). `ColMAD` still outperforms `CopMAD` even with identical debaters, while all `MAD` methods underperform the single-agent baseline (equivalent to Voting when debaters agree), demonstrating the necessity of using heterogeneous LLMs for effective error detection.

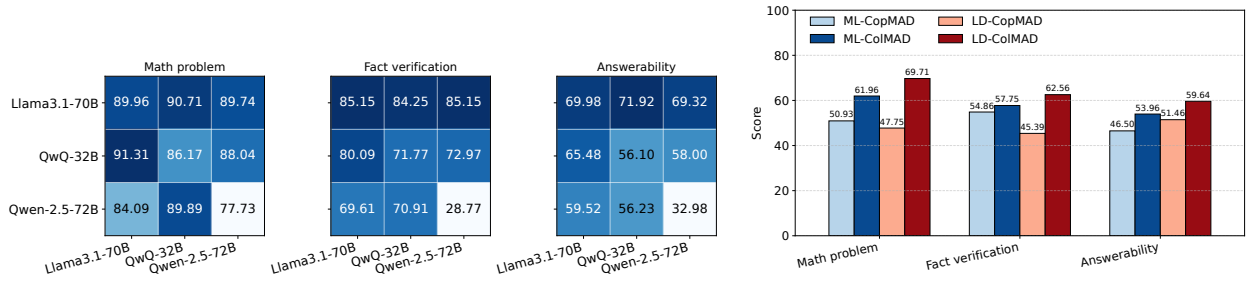
Debater#1	Debater#2	Judge	Math Problem F1	Math Problem F2	Fact Verification F1	Fact Verification F2	Answerability F1	Answerability F2	Average F1	Average F2
GPT4o-mini	GPT4o-mini	GPT4o-mini	**76.58**	**87.99**	**72.82**	**78.89**	**70.00**	**75.92**	**73.13**	**80.93**
			70.90	74.44	66.28	66.74	48.23	42.29	61.80	61.16
			78.70	89.10	75.76	82.78	70.10	74.73	74.85	82.20
GPT4o-mini	GPT4o-mini	Mistral-7B-v0.3	**73.68**	**81.91**	**65.22**	**68.34**	**68.75**	**72.85**	**69.22**	**74.37**
			61.35	58.96	54.88	53.70	53.15	47.03	56.46	53.23
			78.70	89.10	75.76	82.78	70.10	74.73	74.85	82.20

Rows within each block, top to bottom: `ColMAD`, `CopMAD`, single-agent (= Voting when both debaters are the same).

Table 10. Averaged token cost of debating transcripts across `MAD` protocols, with `Llama3.1-70B` and `GPT4o-mini` as debaters. Numbers after $\pm$ are standard deviations. Although `ColMAD` consumes somewhat more tokens than other protocols, the improvement in F2 is substantially larger than the $\sim$15% token overhead over `CopMAD`.

Protocol	Math Problem F2	Math Problem Token cost	Fact Verification F2	Fact Verification Token cost	Answerability F2	Answerability Token cost
`ColMAD`	**90.06**	13259.15 ± 971.28	**88.98**	13327.00 ± 1033.25	**79.74**	13323.00 ± 1086.06
`CopMAD`	88.91	11449.09 ± 524.24	82.43	11484.94 ± 538.15	69.32	11401.81 ± 714.19
SoM	89.55	9461.82 ± 4225.49	79.37	6642.09 ± 1061.34	60.14	7076.89 ± 4165.38
MP	75.79	9690.29 ± 2224.01	75.00	9344.22 ± 712.82	55.56	9471.56 ± 2136.85

Explanation alignment. Fig. 8(b) shows `ColMAD` yields more human-aligned explanations than `CopMAD`, which is critical for scalable oversight. Fine-grained alignment results are given in Appendix F.5.



(a) Combination of different LLMs

(b) Explanation alignment

Figure 8. (a) F2 scores of `ColMAD` under different combinations of LLMs, where the diagonal line shows the single-agent performance. (b) Rate of alignment to the ground-truth explanations given by `CopMAD` and `ColMAD`, where "ML-" refers to the combination of `GPT4o-mini` and `Llama3.1-70B`, and "LD-" refers to the combination of `Llama3.1-70B` and `DeepSeek-R1`. `ColMAD` yields more reasonable explanations than `CopMAD`.

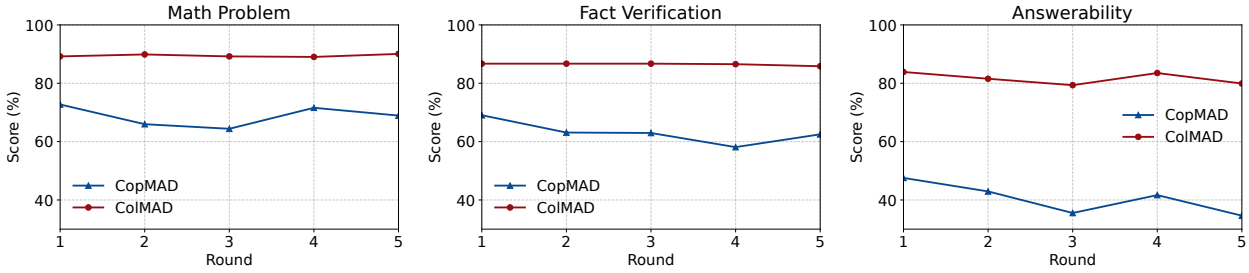
Sensitivity to debate rounds. In Fig. 9, `ColMAD` is generically robust to the number of debate rounds, while `CopMAD` degrades with more rounds. The results are also consistent with babbling accumulation under competitive incentives, where the debate messages become more uninformative with more rounds.

Table 11. Fine-grained coverage (%) of human-annotated error categories in math problem generation, with `Llama3.1-70B` and `DeepSeek-R1` as debaters. `ColMAD` consistently identifies the same mistakes pointed out by humans better than the single-agent and `CopMAD` alternatives.

Error category	Llama3.1-70B	DeepSeek-R1	`CopMAD`	`ColMAD`
Specific integer format	17.60	38.20	38.20	**44.10**
Specific phase requirement	50.00	58.30	58.30	**58.30**
Ambiguous or unanswerable	30.00	40.00	30.00	**40.00**
Missing required content	14.29	14.29	0.00	**28.57**
Calculation error	0.00	50.00	0.00	**50.00**

Table 12. Mean $\pm$ std and 95% confidence interval of F2 scores over 10 runs (temperature $= 1$, seeds 1–10) with `GPT4o-mini` and `Llama3.1-70B` as debaters. `ColMAD` is the only method that achieves non-overlapping 95% CIs over the single-agent methods on fact verification and average F2.

Method	Math Problem F2		Fact Verification F2		Answerability F2		Avg. F2	
	mean±std	95% CI	mean±std	95% CI	mean±std	95% CI	mean±std	95% CI
GPT4o-mini	88.82 ± 1.18	[87.93, 89.71]	78.71 ± 2.11	[77.12, 80.30]	75.06 ± 3.67	[72.29, 77.83]	80.86 ± 1.54	[79.71, 82.02]
Llama3.1-70B	87.11 ± 1.29	[86.14, 88.08]	82.31 ± 1.64	[81.07, 83.54]	62.52 ± 3.91	[59.57, 65.46]	77.31 ± 1.36	[76.29, 78.33]
`CopMAD`	77.00 ± 3.53	[74.34, 79.65]	63.77 ± 3.84	[60.87, 66.67]	46.14 ± 3.93	[43.17, 49.10]	62.30 ± 2.01	[60.78, 63.82]
Voting	88.60 ± 1.13	[87.74, 89.45]	81.15 ± 1.36	[80.12, 82.17]	68.40 ± 3.58	[65.70, 71.10]	79.38 ± 1.33	[78.38, 80.38]
`ColMAD`	$\mathbf{89.53 \pm 0.80}$	$[\mathbf{88.93}, \mathbf{90.13}]$	$\mathbf{85.40 \pm 1.34}$	$[\mathbf{84.39}, \mathbf{86.41}]$	$\mathbf{75.20 \pm 3.81}$	$[\mathbf{72.33}, \mathbf{78.07}]$	$\mathbf{83.38 \pm 1.11}$	$[\mathbf{82.54}, \mathbf{84.21}]$



(a) Math problem (b) Fact verification (c) Answerability

Figure 9. `ColMAD` is generically robust to the number of rounds for debate.