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**[3,2], **Masashi Sugiyama**[2,4]

[1]The Chinese University of Hong Kong, [2]RIKEN Center for Advanced Intelligence Project,
[3]Hong Kong Baptist University, [4]The University of Tokyo
{yqchen,jcheng}@cse.cuhk.edu.hk, gang.niu.ml@gmail.com
bhanml@comp.hkbu.edu.hk, sugi@k.u-tokyo.ac.jp

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 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 consistently outperforms previous MAD protocols under multi-seed evaluation across heterogeneous LLM pairs. 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. The code is available at https://github.com/LFhase/ColMAD.

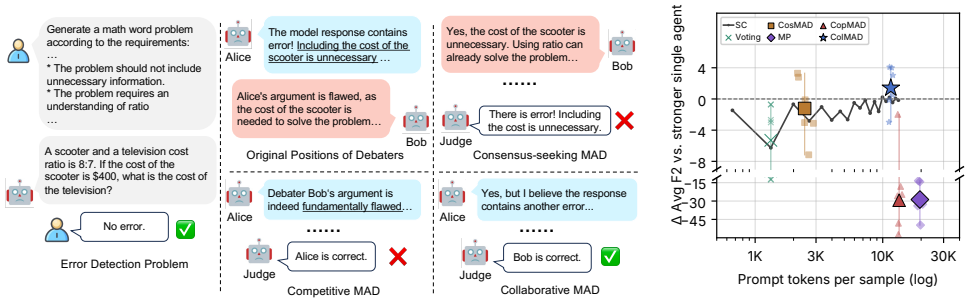
1 Introduction

Along with the success of Large language models (LLMs) in solving tasks at various complexity levels [6, 17, 20], it has also drawn huge attention from the community to explore whether LLMs can exhibit collective intelligence and 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 [7, 12, 16]. MAD has shown success in improving the reasoning of LLMs [15], or even supervising powerful LLMs beyond human intelligence [18]. It further motivates more sophisticated MAD approaches in improving the communication among the agents [9, 40], and debating protocols [11, 24, 25, 28].

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 sig-

*Work done during an internship at RIKEN Center for Advanced Intelligence Project.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Illustration of MAD on error detection

(b) Cost versus gain

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) Change in average F2 over the stronger single agent against prompt tokens per sample on the six pairs of LLMs from Table 1. ColMAD is the only protocol whose mean is above the single agent and its worst pair is the mildest; CosMAD is cheap but below zero on average; CopMAD loses 14 to 45 points on four of the six pairs and MP 24 and 40 on the two pairs with a reasoning model.

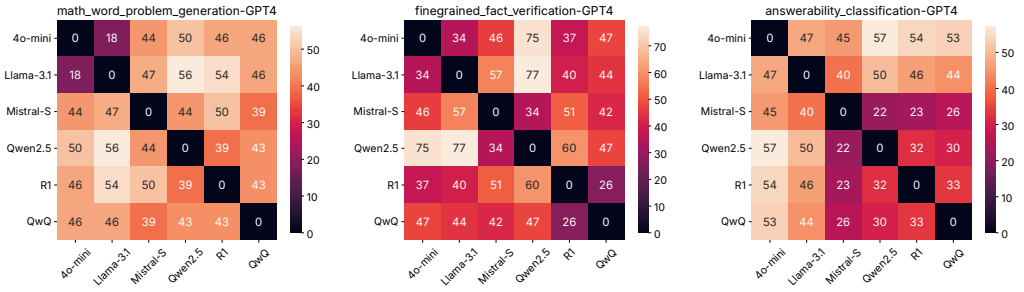
nificantly larger token costs of MAD [27, 32, 36, 39, 42]. When SA is scaled with self-consistency [37] under comparable token budgets, the gap widens further, raising a natural question:

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 [7, 18, 24, 25, 28]; (ii) consensus-seeking MAD (CosMAD): agents may hold different positions and aim to reach to a consensus [9, 10, 15, 28, 40]. We examine the effectiveness of MAD through the task of error detection ReaLMistake [22], where agents need to find whether there is a mistake in a given LLM response. Since LLMs differ in their knowledge and error tendencies [26], different LLMs may provide *complementary signals for error detection*, and jointly considering perspectives from different LLMs becomes more critical. Although MAD has a great potential in error detection (see Fig. 2), CopMAD falls well below a single agent (see Fig. 3), and CosMAD improves over the single agents only when the two debaters have comparable capability.

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 relate debate hacking to *goal misalignment* in the protocol design: the incentive to *win the game* in CopMAD admits equilibria where the messages carry no information about the answer, so competitive debate guarantees no improvement in the worst case, while the incentive to *agree with each other* in CosMAD discards informative disagreements. 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, so that the judge can make a more informative decision from the transcripts (Proposition 2.8). Empirically, on the three error-detection tasks of ReaLMistake [22] with six heterogeneous LLM pairs, ColMAD consistently outperforms both CopMAD and CosMAD (Sec. 4). Multi-seed evaluations also confirm the statistical significance of ColMAD over CopMAD, CosMAD and SA. Extended comparisons on mathematical reasoning and safety alignment [39] are



(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.

consistent with this picture, and under the same token budgets `ColMAD` also outperforms `SA` scaled with self-consistency and with voting: proper protocol design is what unlocks 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 $M^{(t-1)}$.*

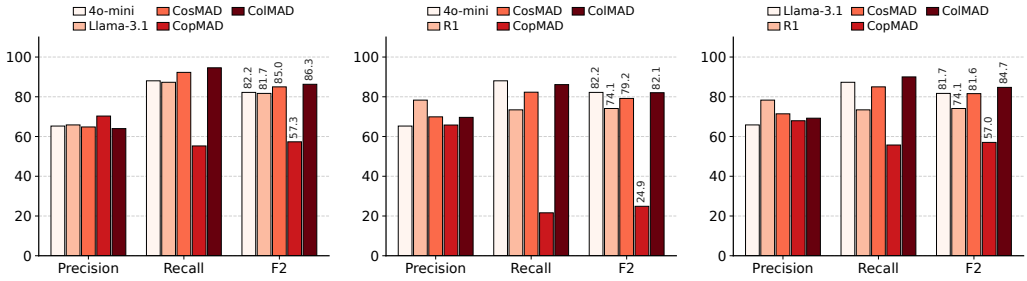
We examine the potential of `MAD` in the task of error detection via the `ReaLMistake` benchmark [22]. It contains 3 objective error detection tasks: (i) math word problem generation (Math problem) that requires LLMs to generate math problems satisfying requirements; (ii) fine-grained fact verification (Fact verification) that requires LLMs to verify claims given fine-grained evidence; (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 [26] and single LLM can hardly tell the errors in LLM responses when without no reliable external feedback [23], 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 [18, 24, 25].

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 debating for the *correct* answer, the judge can be convinced to take the correct answer whenever either agent 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

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

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 six LLMs used in our experiments in Fig. 2. Different LLMs tend not to make the same mistakes: across all pairs and tasks, the oracle collaboration reduces the errors by 18% to 77%. This demonstrates the substantial untapped potential of ensembling opinions from different LLMs.



(a) 4o-mini + Llama-3.1 (b) 4o-mini + R1 (c) Llama-3.1 + R1

Figure 3: Precision, recall and F2 of SA and MAD protocols on three debater pairs. Debater#1 is listed first and also serves as the judge where applicable. CopMAD falls far below the stronger SA on all pairs. CosMAD exceeds both SA in (a) and does not improve over the stronger one in (b) and (c). ColMAD achieves the highest F2 in (a) and (c) and stays at the level of the stronger SA in (b).

Appendix D extends the analysis to more LLMs and to the responses of Llama-2. 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: CopMAD (i.e., Competitive MAD) where LLM agents competitively debate with each other to convince the judge of the respective assigned answers [25]; and CosMAD (i.e., Consensus-seeking MAD) where LLM agents seek consensus [15]. In Fig. 3, we conduct an initial experiment with CopMAD and CosMAD adapted from previous CopMAD [24, 25] and CosMAD [15], respectively. Specifically, we consider three debater pairs formed by GPT4o-mini (4o-mini in short), Llama-3.1-70B (Llama-3.1 in short) and DeepSeek-R1 (R1 in short). We report the precision, recall, and F2 score, which is an instantiation of the F_β score [3]: $F_\beta = (1+\beta^2) \cdot \text{precision} \cdot \text{recall} / (\beta^2 \cdot \text{precision} + \text{recall})$, with β as 2. We focus on the F2 score as it emphasizes recall, i.e., whether all errors are detected, which is critical to error detection.

From the results, CopMAD and CosMAD show different failure patterns. CopMAD cuts the recall sharply and falls far below the stronger single agent on all three pairs. CosMAD is a strong baseline when the two debaters have comparable F2, but no longer improves over the stronger single agent when their F2 differs substantially. 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, CopMAD exhibits the failure widely observed in the literature [32, 36, 39, 42], and CosMAD improves over SA only when the debaters have comparable F2, 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 $\Lambda = \Lambda_0(X_0) + l_A(m_A; X_0) + l_B(m_B; X_0)$.*

Intuitively, Assumption 2.2 formalizes that debate can help the judge only when messages carry information beyond X_0, i.e., $I(Y; M \mid X_0) > 0$. Otherwise $\Lambda(X_0, M) = \Lambda_0(X_0)$ almost surely, where $\Lambda(X_0, Z) = \log \frac{P(Y=1|X_0,Z)}{P(Y=0|X_0,Z)}$ denotes the posterior log-odds given (X_0, Z), and the transcript does not change the judge's decision. In CopMAD [24, 25], 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 $\mathcal{E}_{\text{cop}}(J)$ is the set of Nash equilibria of the A-B subgame in convincing J, M_e is the debating transcript under e, and the inner maximum selects the equilibrium that is worst for the judge.*

CopMAD creates a *zero-sum* incentive: each debater benefits from convincing the judge, *regardless of correctness*. Essentially, CopMAD naturally extends the typical cheap talk game [14], where debaters A and B can transmit costless messages to each other and to the judge J. Since each debater's utility depends only on the judge's verdict, which is a function of (X_0, M), the subgame admits a *babbling equilibrium* for any judge, in which $M \perp Y \mid X_0$ and the 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 is a worst-case statement: under the babbling equilibrium, $I(Y; M \mid X_0) = 0$ and the Bayes-optimal judge gains nothing from the transcript, so the competitive incentive alone guarantees no improvement over R_0. This is compatible with Khan et al. [25] and Kenton et al. [24], who report settings where competitive debate is informative and supports scalable oversight, i.e., an informative equilibrium is reached. In practice, real judges are not Bayes-optimal and are susceptible to debate hacking, which can place CopMAD *below* the SA baseline [42]. On error detection, the fabricated evidence and overconfident claims in Fig. 4 are consistent with debaters reaching unfavorable equilibrium. In the meantime, Khan et al. [25] and Kenton et al. [24] report settings where competitive debate is informative and supports scalable oversight, i.e., an informative equilibrium is reached. Proposition 2.4 is compatible with these findings: it shows that such gains are not guaranteed by the competitive incentive. On error detection, the fabricated evidence and overconfident claims in Fig. 4 are consistent with debaters reaching the unfavorable equilibria. Let us then discuss CosMAD [10, 15], where 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 $V_{\text{cos}} = \min_J \min_{e \in \mathcal{E}_{\text{cos}}(J)} P(J(X_0, M_e) \neq Y)$, 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.*

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 $P\big(\Lambda(X_0, \Phi_\cap(M))\, \Lambda(X_0, M) < 0\big) > 0$, i.e., when the disagreements discarded by the consensus process would overturn the judge's decision with positive probability.*

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. To mitigate the issue, we introduce Collaborative Multi-Agent Debate (ColMAD) to encourage both conditions.

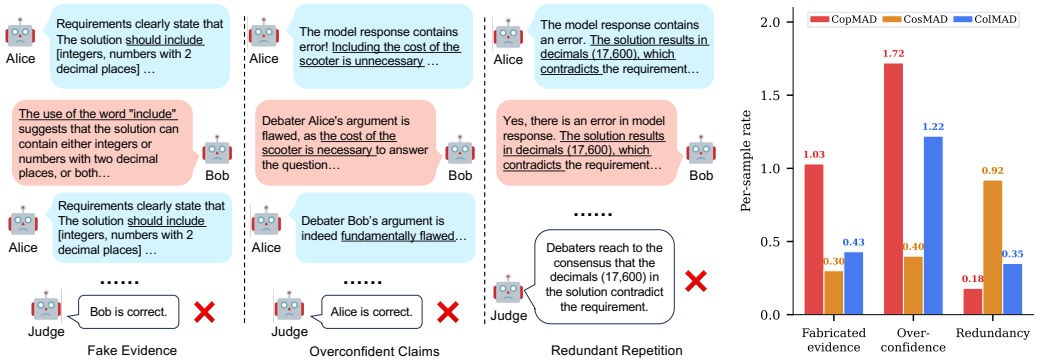
Definition 2.7 (Collaborative MAD)**.** *Debaters* A *and* B *have truth-seeking utilities: $u_A(m_A) = u_B(m_B) = I(Y; m_i \mid X_0, M^{(t-1)})$, i.e., both debaters aim to reduce uncertainty about Y by complementing missing information. The risk is $V_{\text{col}} = \min_J \min_{e \in \mathcal{E}_{\text{col}}(J)} P(J(X_0, M_e) \neq Y)$, where M_e is the debating 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)**.** $V_{\text{col}} \leq V_{\text{cos}} \leq R(X_0) = V_{\text{cop}}$.



(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.

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 [29]. Recently, MAD has gained significant attention as one of the promising approaches to scale up the test-time computation for enhancing reasoning and alignment [15, 19, 24, 25]. 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 [15]. A judge agent can also be incorporated to read the transcripts of the debate and to give a final answer [25]. MAD has demonstrated great potential in improving the reasoning and reducing the hallucinations of LLMs [11, 15, 28, 40]. Liang et al. [28] further assigned personalities to the debating agents to explore the potential answers more efficiently. Yin et al. [40] manually specified diverse roles to agents, and proposed a confidence-based mechanism to reduce the error propagation during reasoning. Chen et al. [11] 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 [19, 24, 25]. Khan et al. [25] showed that LLMs tend to be more persuasive when debating for the correct answer. Kenton et al. [24] 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 [22] as the primary testbed to study the usefulness of MAD. Our taxonomy also has a counterpart in the study of human argumentation Walton and Krabbe [35]: CopMAD resembles a persuasion dialogue, where each party aims to convince the other of its own thesis, while the goal structure of ColMAD resembles an inquiry dialogue, where the participants share the aim of establishing whether a proposition is true.

Different from Walton and Krabbe [35], we study debates among LLM agents, identify the failures of existing MAD protocols through their incentive structures, and derive a proper protocol design.

Pitfalls of Multi-Agent Debate. Recent studies also showed the limitations of MAD [32, 36, 42]. Wang et al. [36] found that SA methods can already perform competitively or outperform MAD when given sufficient information. Smit et al. [32] showed that MAD can underperform SA when scaling SA using methods such as self-consistency [37]. Zhang et al. [42] and Kim et al. [27], Yang et al. [39] provided more evidence to support the findings by Wang et al. [36] and Smit et al. [32]. 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 of when and why MAD succeeds and fails. Similar to Wang et al. [36], Smit et al. [32], Zhang et al. [42], and Yang et al. [39], we find that the vanilla MAD protocol can often lead to degraded performances, i.e., 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 [22], 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 [13], AIME-2024 [2], and Anthropic-Harmful [41], following Yang et al. [39].

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. [24] that demonstrated impressive capabilities in error detection for CopMAD, and SoM [15] for CosMAD. For the implementation of SoM, we follow the practice of Yang et al. [39]. We also include another popular CopMAD scheme, MP [28], that incorporates persona into agents. In addition, we also consider a simple **Voting** baseline, which can be considered the simplest collaborative MAD. The reasoning-task and safety-task experiments in Table 3 also follow the setting of Yang et al. [39].

LLM Backbones. As the original LLMs benchmarked in ReaLMistake [22] are a bit outdated, we incorporate new frontier LLMs, including GPT4o-mini [30], Llama3.1-70B [1], Mistral-Small-2603/Mistral-7B-v0.3 [21], Qwen-2.5-72B [33] as well as the frontier reasoning LLMs including DeepSeek-R1 [17] and QwQ-32B [34]. 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 [3]: $F_\beta = (1+\beta^2) \cdot \text{precision} \cdot \text{recall} / (\beta^2 \cdot \text{precision} + \text{recall})$, 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:

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

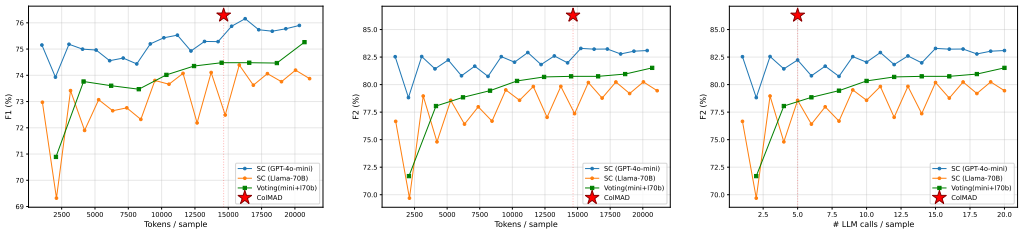
Debater#1	Debater#2	Protocol	Math Problem		Fact Verification		Answerability		Avg. F1	Avg. F2
			F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)	F1 (↑)	F2 (↑)
Human	-	-	90.00	84.91	95.45	95.45	90.48	87.96	89.44	91.98
GPT4o-mini	-	-	78.70	89.10	75.76	82.78	70.10	74.73	74.85	82.20
Llama3.1-70B	-	-	79.26	89.96	76.85	85.15	68.13	69.98	74.75	81.70
Mistral-Small-2603	-	-	76.24	78.05	55.17	50.00	44.62	37.08	58.68	55.04
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	65.50	64.81	65.84	63.70	48.85	40.82	60.06	56.44
		CosMAD	77.83	89.21	76.33	85.50	**74.00**	**80.26**	76.05	84.99
		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	78.03	89.88	77.42	**88.98**	72.91	79.74	76.12	**86.20**
		CopMAD	78.34	88.91	73.79	82.43	68.16	69.32	73.43	80.22
		CosMAD	78.73	90.25	77.73	87.98	72.36	78.26	76.27	85.50
		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-Small-2603	ColMAD	80.00	89.17	76.14	82.96	65.52	65.52	73.89	79.22
		CopMAD	52.94	45.34	46.88	39.16	24.07	17.62	41.30	34.04
		CosMAD	77.45	84.95	75.51	82.04	59.17	58.14	70.71	75.04
		MP	68.21	67.97	63.41	62.05	38.33	30.18	56.65	53.40
		Voting	80.00	86.77	65.90	66.59	59.63	56.87	68.51	70.08
Llama3.1-70B	Mistral-Small-2603	ColMAD	79.07	89.29	79.43	**89.44**	66.30	67.87	74.93	82.20
		CopMAD	70.52	70.28	66.31	70.14	56.16	50.37	64.33	63.60
		CosMAD	79.21	86.39	76.47	84.97	64.37	64.37	73.35	78.58
		MP	49.21	40.05	61.64	56.11	53.90	47.26	54.92	47.81
		Voting	76.62	83.33	68.97	69.93	58.06	54.09	67.88	69.12
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	83.42	90.22	78.57	85.37	63.47	61.92	75.15	79.17
		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	**83.84**	**90.41**	78.39	85.90	69.41	68.45	**77.21**	81.59
		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

Table 2: F2 scores for GPT-4 responses over 10 seeds at temperature 1. The best result is in bold and the second best is underlined. * marks the columns in which ColMAD outperforms CosMAD with $p < 0.05$ (paired t test), and ColMAD outperforms CopMAD in every column ($p < 10^{-4}$).

Debater#1	Debater#2	Protocol	Math Problem	Fact Verification	Answerability	Avg. F2
GPT4o-mini	-	-	89.14 ± 0.73	78.71 ± 2.22	75.06 ± 3.87	80.97 ± 1.53
Llama3.1-70B	-	-	87.11 ± 1.36	82.31 ± 1.72	62.52 ± 4.12	77.31 ± 1.43
Mistral-Small-2603	-	-	78.42 ± 6.27	61.30 ± 3.46	47.40 ± 7.86	62.37 ± 4.68
GPT4o-mini	Llama3.1-70B	ColMAD	$\underline{89.47 \pm 0.90}$	$\mathbf{85.08 \pm 1.65}$	$\mathbf{76.05 \pm 3.94}^*$	$\mathbf{83.53 \pm 1.57}^*$
		CopMAD	76.75 ± 3.09	63.78 ± 4.64	45.69 ± 4.01	62.07 ± 1.54
		CosMAD	$\mathbf{89.64 \pm 0.73}$	$\underline{83.43 \pm 2.16}$	$\underline{73.35 \pm 4.63}$	$\underline{82.14 \pm 1.88}$
Llama3.1-70B	GPT4o-mini	ColMAD	$\mathbf{89.23 \pm 1.12}$	$\mathbf{86.52 \pm 1.57}^*$	$\mathbf{73.48 \pm 2.76}$	$\mathbf{83.08 \pm 1.53}^*$
		CopMAD	82.44 ± 1.96	74.42 ± 2.99	56.16 ± 4.46	71.01 ± 1.07
		CosMAD	$\underline{89.15 \pm 0.76}$	$\underline{85.28 \pm 1.05}$	$\underline{72.87 \pm 4.01}$	$\underline{82.44 \pm 1.60}$
GPT4o-mini	Mistral-Small-2603	ColMAD	$\mathbf{88.06 \pm 1.15}^*$	$\mathbf{81.08 \pm 2.17}^*$	$\mathbf{69.16 \pm 4.60}^*$	$\mathbf{79.43 \pm 1.51}^*$
		CopMAD	54.30 ± 5.30	47.98 ± 3.93	26.43 ± 5.50	42.90 ± 2.17
		CosMAD	$\underline{86.65 \pm 2.19}$	$\underline{79.30 \pm 1.86}$	$\underline{64.62 \pm 5.74}$	$\underline{76.86 \pm 2.67}$
Llama3.1-70B	Mistral-Small-2603	ColMAD	$\mathbf{87.80 \pm 2.33}$	$\mathbf{85.52 \pm 1.28}^*$	$\mathbf{66.01 \pm 4.37}^*$	$\mathbf{79.78 \pm 1.97}^*$
		CopMAD	65.07 ± 3.94	69.45 ± 3.09	47.77 ± 4.86	60.77 ± 2.27
		CosMAD	$\underline{86.58 \pm 2.58}$	$\underline{81.90 \pm 2.39}$	$\underline{63.15 \pm 3.85}$	$\underline{77.21 \pm 1.90}$

Finding 1: Single agents still struggle with error detection, and existing MAD protocols do not reliably improve on them. 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. 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 improves over both single agents on the two comparable pairs, but falls below the best single agent whenever the other debater is weaker.

Finding 2: ColMAD consistently outperforms all MAD baselines and improves over SA on comparable-capability pairs. Under multi-seed evaluation (Table 2), ColMAD outperforms CosMAD on all four pairs and CopMAD by 12 to 37 points. In Table 1, ColMAD outperforms CopMAD, SoM, and

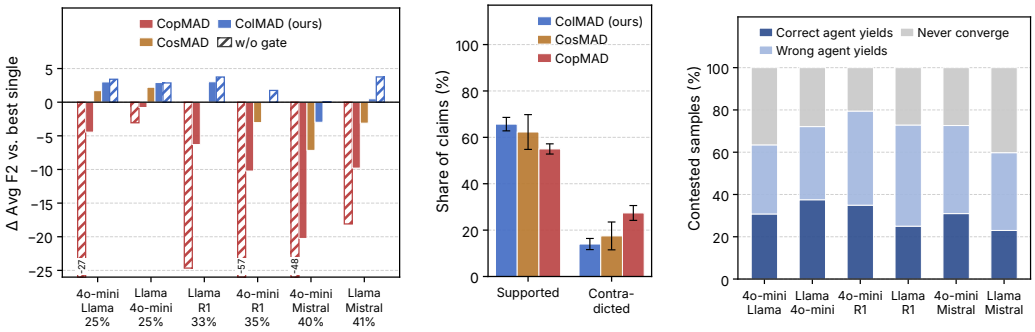


(a) F1 under different token costs (b) F2 under different token costs (c) F2 under different LLM calls

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.

MP on all six pairs in average F2 with zero exceptions. Against the stronger single agent, ColMAD improves average F2 by 2.6 and 2.1 points over 10 seeds on the two comparable pairs, with the gain concentrated on the task where each agent trails its partner. With a much weaker partner the gain shrinks. 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 3: ColMAD outperforms single-agent scaling and scaled voting under matched compute. A natural concern is whether ColMAD's gains come from using more tokens. As shown in Fig. 4.2, we compare ColMAD with SA using self-consistency (SC) and with Voting scaled by sampling both agents m times. One can find that, SC saturates quickly for both models, with F2 oscillating regardless of budget, and scaled Voting saturates at about 82 average F2 by 28 calls. At matched tokens (∼14.7K), ColMAD outperforms SC@14 for GPT4o-mini and Llama. Under the same LLM calls, the gap is larger still, and ColMAD outperforms Voting even at 28 calls. This confirms that ColMAD's advantage comes from effective exploitation of the *cross-model diversity* instead of compute scaling.



(a) Gain vs. disagreement rate and gate (b) Claim ratings (c) How CosMAD resolves disagreements

Figure 6: (a) Change in average F2 over the per-task best single agent. (b) Share of claims in the transcripts of contested samples rated supported or contradicted by an independent rater (mean ± sd); (c) How CosMAD resolves each contested sample of the six pairs.

Finding 4: ColMAD is the most stable protocol as disagreement grows, and elicits the most supported claims on contested samples. Algorithm 1 skips the debate whenever the two initial answers agree, while CopMAD debates every sample as originally specified [24, 25]. Adding the gate to CopMAD raises its performance, yet gated CopMAD still falls below ColMAD on every pair, and below ColMAD without the gate as well. As the initial disagreement rate grows from 25% to 40%, the performances of CopMAD and CosMAD vary more than ColMAD. On the contested samples, ColMAD has the largest share of supported claims and the most decisive claims per sample, i.e., supported claims that settle the verdict by themselves, while CopMAD has the largest share of contradicted claims, consistent with Sec. 2.2. CosMAD resolves the contested samples largely independently of correctness, consistent with Proposition 2.6.

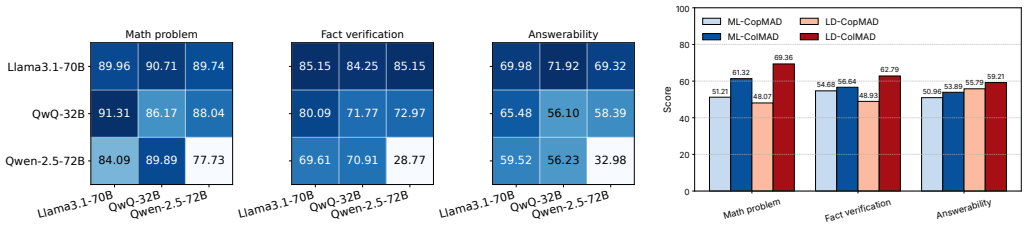
4.3 Ablation Studies and Analysis

In Table 3, we extend to reasoning and safety benchmarks following Yang et al. [39], using Qwen2.5-7B and Llama3.1-8B. ColMAD improves over CosMAD and the best SA on the harder AIME-2024 at

Table 3: Performance of different test-time scaling methods under different LLM call budgets.

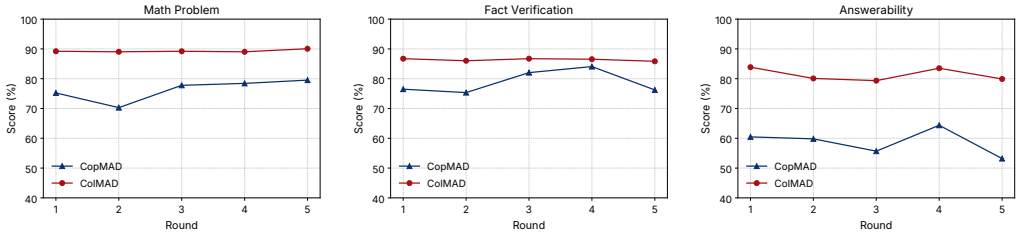
Debater#1	Debater#2	Protocol	GSM8K (ACC ↑)			AIME-2024 (ACC ↑)			Anthropic-Harmful (ASR ↓)		
			4	8	16	4	8	16	2	4	8
Qwen2.5-7B	-	-	**90.40**	**91.40**	**90.80**	13.33	13.33	13.33	14.59	18.92	31.69
Llama3.1-8B	-	-	88.80	89.60	89.40	3.33	6.67	6.67	14.32	9.93	15.34
Qwen2.5-7B	Llama3.1-8B	ColMAD	90.00	90.20	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

the highest call budget and dramatically reduces the attack success rate on Anthropic-Harmful to near 0% with cross-model debate, suggesting strong potential for safety-critical oversight. On saturated tasks like GSM8K, improvements saturate, consistent with Proposition 2.8: little room remains for complementary information.



(a) Combination of different LLMs (b) Explanation alignment

Figure 7: (a) shows the results in F2 scores of ColMAD performance under different combinations of LLMs, where the diagonal line shows the single-agent performance. (b) shows the 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.



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

Figure 8: ColMAD is generically robust to the number of rounds for debate.

Judge robustness. Different combinations of LLMs as judges are shown in Fig. 7(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.

Explanation alignment. Fig. 7(b) shows ColMAD yields more human-aligned explanations than CopMAD, critical for scalable oversight. Fine-grained alignment results are given in Appendix F.5.

Sensitivity to debate rounds. In Fig. 8, ColMAD is 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.

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 on error detection, together with extended evaluations on reasoning and safety tasks, show that MAD is not inherently ineffective: with heterogeneous debaters and a proper protocol design, MAD can improve over SA methods.

Acknowledgments and Disclosure of Funding

We thank the reviewers and the area chair for their constructive comments, which directly shaped the additional analyses in this paper. MS was supported by JST ASPIRE Grant Number JPMJAP2405. BH was supported by NSFC General Program No. 62376235, RGC General Research Fund No. 12200725, RGC Research Fellow Scheme No. RFS2627-2S02, RGC Young Collaborative Research Grant No. C2005-24Y, RGC General Research Fund No. 12202026, NSFC Major Research Plan No. 92570109, RIKEN Collaborative Research Fund, Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme, HKBU Faculty Niche Research Areas No. RC-FNRA-IG/25-26/SCI/05, and HKBU CSD Departmental Incentive Scheme.

References

[1] M. AI. 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 7)

[2] 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 7)

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval: The concepts and technology behind search*. Addison-Wesley Publishing Company, USA, 2nd edition, 2011. ISBN 9780321416919. (Cited on pages 4 and 7)

[4] J. Brown-Cohen, G. Irving, and G. Piliouras. Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125*, 2023. (Cited on page 18)

[5] J. Brown-Cohen, G. Irving, and G. Piliouras. Avoiding obfuscation with prover-estimator debate. *arXiv preprint arXiv:2506.13609*, 2025. (Cited on page 18)

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

[7] M. D. Buhl, J. Pfau, B. Hilton, and G. Irving. An alignment safety case sketch based on debate. *arXiv preprint arXiv:2505.03989*, 2025. (Cited on pages 1, 2 and 18)

[8] M. V. Carro, D. A. Mester, F. Nieto, O. A. Stanchi, G. E. Bergman, M. Leiva, L. N. F. Gangi, E. Sprejer, F. G. Selasco, J. G. Corvalan, M. V. Martinez, and G. Simari. 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 18)

[9] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023. (Cited on pages 1, 2 and 18)

[10] J. Chen, S. Saha, and M. Bansal. 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, 2024. (Cited on pages 2, 5 and 18)

[11] P. Chen, S. Zhang, and B. Han. CoMM: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1720–1738, 2024. (Cited on pages 1, 6 and 18)

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

[13] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint*, arXiv:2110.14168, 2021. (Cited on page 7)

[14] V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982. (Cited on pages 5 and 16)

[15] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. (Cited on pages 1, 2, 4, 5, 6, 7, 15 and 18)

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

[17] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. (Cited on pages 1 and 7)

[18] G. Irving, P. Christiano, and D. Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. (Cited on pages 1, 2, 3 and 18)

[19] G. Irving, P. F. Christiano, and D. Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. (Cited on page 6)

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

[21] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts. *arXiv preprint*, arXiv:2401.04088, 2024. (Cited on page 7)

[22] R. Kamoi, S. S. S. Das, R. Lou, J. J. Ahn, Y. Zhao, X. Lu, N. Zhang, Y. Zhang, R. H. Zhang, S. R. Vummanthala, S. Dave, S. Qin, A. Cohan, W. Yin, and R. Zhang. Evaluating llms at detecting errors in LLM responses. *arXiv preprint arXiv:2404.03602*, 2024. (Cited on pages 2, 3, 6, 7 and 23)

[23] R. Kamoi, Y. Zhang, N. Zhang, J. Han, and R. Zhang. 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, 2024. (Cited on page 3)

[24] Z. Kenton, N. Y. Siegel, J. Kramar, J. Brown-Cohen, S. Albanie, J. Bulian, R. Agarwal, D. Lindner, Y. Tang, N. Goodman, and R. Shah. 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, 2, 3, 4, 5, 6, 7, 9, 18 and 20)

[25] A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, and E. Perez. Debating with more persuasive LLMs leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024. (Cited on pages 1, 2, 3, 4, 5, 6, 9, 15 and 18)

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

[27] Y. Kim, K. Gu, C. Park, C. Park, S. Schmidgall, A. A. Heydari, Y. Yan, Z. Zhang, Y. Zhuang, Y. Liu, M. Malhotra, P. P. Liang, H. W. Park, Y. Yang, X. Xu, Y. Du, S. Patel, T. Althoff, D. McDuff, and X. Liu. Capable language models can outgrow the benefits of collaboration. *Nature Machine Intelligence*, 8:1157–1172, 2026. doi: 10.1038/s42256-026-01268-y. (Cited on pages 2 and 7)

[28] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023. (Cited on pages 1, 2, 6, 7 and 18)

[29] M. Minsky. The society of mind. *The Personalist Forum*, 1987. (Cited on page 6)

[30] OpenAI. Gpt-4o mini technical report. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. (Cited on page 7)

[31] S. Rahman, S. Issaka, A. Suvarna, G. Liu, J. Shiffer, J. Lee, M. R. Parvez, H. Palangi, S. Feng, N. Peng, Y. Choi, J. Michael, L. Jiang, and S. Gabriel. AI debate aids assessment of controversial claims. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. (Cited on page 18)

[32] A. P. Smit, P. Duckworth, N. Grinsztajn, K. ab Tessera, T. D. Barrett, and A. Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*, 2024. (Cited on pages 2, 4, 7 and 18)

[33] Q. Team. Qwen2.5 technical report. *arXiv preprint*, arXiv:2412.15115, 2024. (Cited on page 7)

[34] Q. Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/. (Cited on page 7)

[35] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY, 1995. (Cited on pages 6 and 7)

[36] Q. Wang, Z. Wang, Y. Su, H. Tong, and Y. Song. 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)*, pages 6106–6131, 2024. (Cited on pages 2, 4, 7 and 18)

[37] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. (Cited on pages 2 and 7)

[38] A. Wynn, H. Satija, and G. K. Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025. (Cited on page 18)

[39] Y. Yang, E. Yi, J. Ko, K. Lee, Z. Jin, and S. Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*, 2025. (Cited on pages 2, 4, 7, 9 and 18)

[40] Z. Yin, Q. Sun, C. Chang, Q. Guo, J. Dai, X. Huang, and X. Qiu. 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*, pages 15135–15153, 2023. (Cited on pages 1, 2, 6 and 18)

[41] Y. Zeng, Y. Wu, X. Zhang, H. Wang, and Q. Wu. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*, 2024. (Cited on page 7)

[42] H. Zhang, Z. Cui, X. Wang, Q. Zhang, Z. Wang, D. Wu, and S. Hu. If multi-agent debate is the answer, what is the question? *arXiv preprint arXiv:2502.08788*, 2025. (Cited on pages 2, 4, 5, 7 and 18)

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 4: 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.
$\|\ell_i\| \le L_i$	Bounded manipulability / persuasion budget for debater i.
$\Lambda(x, m_A, m_B)$	Total LLR used by the judge: $\Lambda_0(x) + \ell_A(m_A; x) + \ell_B(m_B; x) + \log \frac{\pi}{1-\pi}$.
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 [25]; and CosMAD (i.e., Consensus-seeking MAD) where LLM agents seek consensus [15].

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*

$V_{\cos} = \min_J \min_{e \in \mathcal{E}_{\cos}(J)} P(J(X_0, M_e) \neq Y)$, *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.*

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 and Sobel [14]. 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 $V_{\text{col}} \leq R(X_0, \Phi_\cap(M)) \leq R(X_0) = V_{\text{cop}}$. □

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.

Table 5: 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.

Table 6: 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. [24]
Khan et al. [25]	Competitive MAD	Khan et al. [25]
Irving et al. [18]	Competitive MAD	Irving et al. [18]
Brown-Cohen et al. [5]	Competitive MAD	Brown-Cohen et al. [5]
Buhl et al. [7]	Competitive MAD	Buhl et al. [7]
Brown-Cohen et al. [4]	Competitive MAD	Brown-Cohen et al. [4]
Rahman et al. [31]	Competitive MAD	Rahman et al. [31]
Carro et al. [8]	Competitive MAD	Carro et al. [8]
MP	Competitive MAD with persona	Liang et al. [28]
SoM	Consensus-seeking MAD (CosMAD)	Du et al. [15]
Reconcile	Consensus-seeking MAD (CosMAD)	Chen et al. [10]
CoMM	Consensus-seeking MAD with persona	Chen et al. [11]
EoT	Communication protocol for MAD	Yin et al. [40]
ChatEval	Application of MAD (evaluation)	Chan et al. [9]
Wang et al. [36]	Benchmarking of MAD	Wang et al. [36]
Smit et al. [32]	Benchmarking of MAD	Smit et al. [32]
Wynn et al. [38]	Benchmarking of CosMAD	Wynn et al. [38]
Zhang et al. [42]	Benchmarking of MAD	Zhang et al. [42]
Yang et al. [39]	Benchmarking of MAD	Yang et al. [39]

B.7 Theory-to-implementation mapping

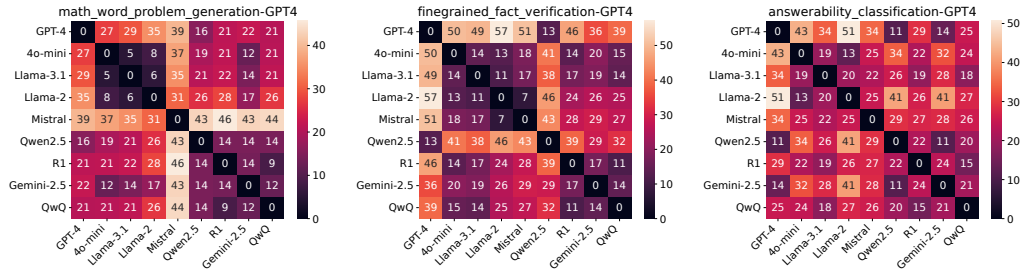
Table 5 maps the theoretical condition that ColMAD requires for its advantage over CopMAD (Proposition 2.8) to the specific prompt components used in our implementation.

C Categorization of existing MAD approaches

We summarize the existing MAD literature in Table 6, 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.

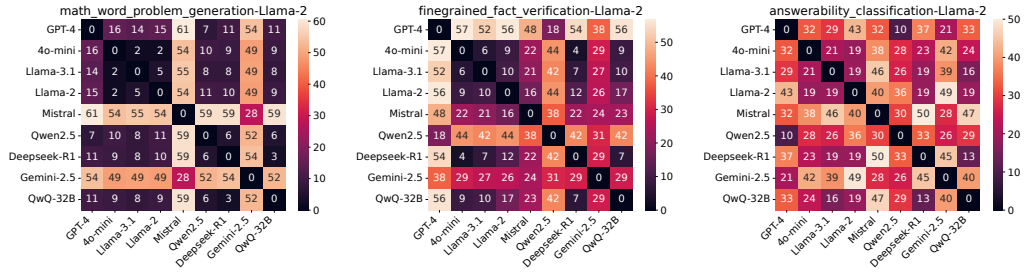
D More Results on Potential of Multi-Agent Collaboration

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



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

Figure 9: 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 10: 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. [24]; (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 7: Statistics of `ReaLMistake` benchmark [22].

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 8. 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 8, 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 9, 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 10. `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.

Table 8: 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 9: 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

F.4 Token cost across MAD protocols

Table 11 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 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. 7(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 12 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.

Table 10: 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 11: 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

Table 12: 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**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our contribution to understanding when and why `MAD` fails, and thereby providing a new `MAD` protocol based on our analysis.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a discussion in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide necessary details in the main text and full details in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full details in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will provide an anonymous link to our code during the discussion phase and open-source the code when published.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the necessary details in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide statistical significance in Table 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Most of our experiments are using API calls, while experiments in Table 3 are run with A100 gpus.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We ensure our paper follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a discussion at the begining of the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work provides new insights for AI safety, and to the best of our knowledge, we are not aware of potential risks caused by our work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite proper sources for data and methods used in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe the use of LLM at the begining of the Appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.