

Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate

Andrea Wynn^{*1} Harsh Satija^{*2} Gillian Hadfield¹³²

Abstract

While multi-agent debate has been proposed as a promising strategy for improving AI reasoning ability, we find that debate can sometimes be harmful rather than helpful. The prior work has exclusively focused on debates within homogeneous groups of agents, whereas we explore how diversity in model capabilities influences the dynamics and outcomes of multi-agent interactions. Through a series of experiments, we demonstrate that debate can lead to a decrease in accuracy over time — even in settings where stronger (i.e., more capable) models outnumber their weaker counterparts. Our analysis reveals that models frequently shift from correct to incorrect answers in response to peer reasoning, favoring agreement over challenging flawed reasoning. These results highlight important failure modes in the exchange of reasons during multi-agent debate, suggesting that naive applications of debate may cause performance degradation when agents are neither incentivized nor adequately equipped to resist persuasive but incorrect reasoning.

1. Introduction

Large Language Model (LLM) agents have demonstrated remarkable problem-solving abilities across a wide array of complex reasoning tasks (Brown et al., 2020). Recently, a new line of research on interactive reasoning among multiple LLMs through debate has promoted the multi-agent debate framework as a promising approach to enhancing the reasoning and decision-making capabilities of LLM agents (Du et al., 2023; Chan et al., 2023; Liang et al., 2023; Khan et al., 2024). Various forms of multi-agent debate have been shown to improve performance on multiple arith-

metic and strategic reasoning benchmarks (Du et al., 2023; Subramaniam et al., 2025), produce more truthful answers and evaluations (Chan et al., 2023; Khan et al., 2024), and enhance tasks such as machine translation (Liang et al., 2023) and negotiation (Fu et al., 2023). The core concept of these studies is that by engaging LLM agents through structured argumentation or discourse, we can facilitate the exchange of reasoning among different agents and guide them toward more accurate answers. However, most of these techniques incur significant computational overhead, as they are inference or test-time methods that rely on the strong zero-shot or few-shot capabilities of LLMs. Intuitively, greater exchanges of reasoning should lead to better decisions—allowing multiple agents to challenge flawed reasoning, highlight overlooked details, and reduce individual biases. But is this always the case?

In this work, we show that the benefits of multi-agent debate are not as universal as commonly assumed. Through a series of empirical studies, we show that multi-agent debate can sometimes degrade performance, leading to worse final answers than those generated by a single agent acting alone. These failures are not rare edge cases, but arise systematically in settings where agents amplify each other’s errors — agreeing reflexively rather than challenging flawed reasoning. These findings hold true even when there is variation in the abilities of the participating LLM agents. For instance, we discover that introducing a weak or less capable (lower-performing) LLM agent into a debate with a strong or more capable (higher-performing) agent can detrimentally affect the debate outcome, producing results worse than if the agents had not engaged in discussion. The presence of a weaker agent disrupts the performance of the stronger agent. Moreover, in certain cases, the longer a debate continues, the more performance can degrade. In other words: talk isn’t always cheap — and in some cases, it’s actively harmful.

We present a systematic evaluation of multi-agent debate across multiple tasks, showing that debate can sometimes *harm* group performance, particularly with heterogeneous LLM agents engaged in debate. Our findings challenge the prevailing narrative that more discussion between agents is inherently beneficial. Instead, we uncover several key factors that mediate the success or failure of debate, including task type and complexity, agent diversity and capability. In doing so, we offer a nuanced view of when and why debate

^{*}Equal contribution ¹Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA ²Vector Institute ³University of Toronto, Toronto, ON, Canada. Correspondence to: Andrea Wynn <awynn13@jhu.edu>, Harsh Satija <harsh.satija@vectorinstitute.ai>.

helps – and when it hurts. Together, these results suggest that while debate remains a promising tool for improving model reasoning, its success is far from guaranteed.

2. Related Work

Debate and multi-agent reasoning. Multi-agent debate was initially proposed as method for scalable oversight problem where a judge or verifier can interject and elicit hidden contradictions, using structured back-and-forth conversations (Irving et al., 2018; Khan et al., 2024; Michael et al., 2023; Kenton et al., 2024). More recently, another form of multi-agent debate, sometimes referred to as multi-agent deliberation, investigates at leveraging different LLM agents to surface better answers by having them exchange reasoning via iterative discussion (Du et al., 2023; Chan et al., 2023; Liang et al., 2023; Subramaniam et al., 2025). Most of these studies focus on a homogeneous setting where all LLM agents utilize the same underlying language model, finding that this approach enhances accuracy across various Question-Answering (QA) tasks.

Estornell & Liu (2024) examine the theoretical properties and effects of opinion diversity within the debate framework, reporting a “tyranny of the majority” effect. They found that if the majority of agents provide the same answer—regardless of its correctness—minority agents tend to conform, creating an echo chamber effect. Their analysis was limited to a homogeneous setting, whereas we extend our analysis to heterogeneous settings. Our findings reveal that even when the majority consists of stronger models, introducing a weaker model can diminish overall performance, thereby complementing their results.

Some works explores debates between agents of diverse nature or abilities. For instance, Estornell et al. (2025); Subramaniam et al. (2025) propose training LLM agents to debate collaboratively with distinct roles (actors/generators and critics), demonstrating that this approach can surpass previous unsupervised debate setups in reasoning benchmarks. Finally, studies such as Amayuelas et al. (2024) investigate whether the collaborative nature holds if an explicit adversary is introduced into the debate process—where the adversary actively seeks to reduce performance.

These works suggest that, when agents constructively challenge each other, answers can improve. However, other studies caution that debate can fail when agents emphasize persuasion over truth. Agarwal & Khanna (2025) introduce a single-round debate on factual questions (using TruthfulQA) where one agent states a true answer calmly and another delivers a confident, emotional false answer. They show the LLM judge often chooses the persuasive falsehood with high confidence, suggesting that a vivid but incorrect argument can override a correct one. These results high-

light a risk: unless the judge is well-calibrated, debate may amplify bluster, mirroring human misinformation scenarios.

Collaborative multi-agent frameworks. Beyond explicit debates, many recent systems assume collaboration among LLMs improves reasoning (Li et al., 2023; Tran et al., 2025). For instance, Wu et al. (2023) provides a general multi-agent conversation framework: developers can define many agents (assistant, user-proxy, tools, etc.) that autonomously chat to solve complex tasks. These role-based and decentralized systems often yield richer interactions (e.g. multi-turn planning) than a single LLM alone.

Sequential revision: Frameworks that utilize interactive reasoning at inference time in a sequential manner are employed for either self-refinement (Madaan et al., 2023) or self-consistency (Wang et al., 2022). Self-refinement involves iteratively revising or adapting a model’s responses based on previous outputs, prompting the model to intentionally reflect on its existing responses and correct any mistakes (Kamoi et al., 2024). Whereas works based on self-consistency often explicitly run multiple reasoning paths in parallel. For example, Wang et al. (2023) sample numerous independent chain-of-thought answers and select the most common answer. This aggregation of diverse reasoning paths significantly enhances the accuracy of arithmetic and commonsense QA by minimizing uncertainty. He et al. (2025) extend this idea: they spawn multiple “reactive” and “reflection” agent pairs, each exploring a different reasoning path, and then use a separate summarizer to aggregate them. Likewise, Yang et al. (2025) build a decentralized multi-agent planner where each LLM agent maintains its own memory (a hierarchical knowledge graph) and communicates via structured prompts. They find that these collaborative agents reach goals with 60% fewer steps than a lone agent, underscoring the value of structured cooperation.

Collectively, prior work shows multi-agent debate and collaboration can amplify reasoning abilities, but also highlights key failure modes: judges may be fooled by rhetoric and human-like dynamics can bias group outputs. Our work builds on these insights by examining exactly when and why debate among LLMs helps or hurts performance on a variety of tasks.

3. Setting: Multi-agent debate

Let \mathcal{Q} denote the a dataset of questions related to a task, where each question $q \in \mathcal{Q}$ in the task is natural language text. The objective is to generate an answer $a \in \mathcal{A}$ for any given input question $q \in \mathcal{Q}$, where \mathcal{A} is the set of possible answers for that task. We assume that there exists a ground truth answer $a^* \in \mathcal{A}$ for each question $q \in \mathcal{Q}$ which is denoted by $f^{gt} : \mathcal{Q} \rightarrow \mathcal{A}$, i.e., $f^{gt}(q) = a^*$.

Single-Agent Setting: In the single agent setting, each

agent uses an underlying LLM $l : \mathcal{Q} \rightarrow \mathcal{A}$ to generate answer $a \sim l(q)$ where $a \in \mathcal{A}$ denotes the answer generated by the LLM in response to the question q . The main metric of interest is accuracy which is calculated as $\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1}[l(q) = f^{gt}(q)]$.

Usually, various task specific prompts are also given as input to the LLM in addition to the input question for generating an answer. Let \mathcal{P}_T denote the task-specific prompt, then the resulting LLM can be described as $l_{\mathcal{P}_T} : \mathcal{P}_T \times \mathcal{Q} \rightarrow \mathcal{A}$ or $a \sim l(\mathcal{P}_T(q))$.

Multi-Agent Debate: We follow the multi-agent debate framework from (Du et al., 2023; Subramaniam et al., 2025) that involves initially posing a question to a group of LLM agents. After the initial responses are generated, the debate process then iteratively revisits the question for each agent, contextualizing input through the collective responses from previous rounds. Specifically, each agent generates a new response to the question based on its prior response and the summarized responses from the other agents. The final answer is based by majority vote among all the agents.

Formally, we have a group of N agents, each with their own LLM l_i that are presented with a question $q \in \mathcal{Q}$ and tasked with generating an answer $a \in \mathcal{A}$. We use $d(l_1, \dots, l_n) : \mathcal{Q} \rightarrow \mathcal{A}$ to denote the debate procedure that takes input some question and set of LLM agents and generates an answer. The debate procedure runs over multiple rounds, where at any given round t , each agent i uses their underlying LLM l_i to iteratively generate an answer in the following manner:

1. **Starting Round ($t = 1$):** Each agent i generates an response $g_i \in \mathcal{G}$ via l_i with an optional additional prompt, $S : \mathcal{Q} \times \mathcal{P}_{\text{starting}} \rightarrow \mathcal{A}$, or $g_i^1 \sim l_i(\mathcal{P}_{\text{starting}}(q))$. Here \mathcal{G} denotes the set of possible generations from LLM.
2. **Debate Rounds ($t = 2, \dots, T$):** For the subsequent rounds, each agent i is given the question q and responses from the other agents from the previous round. Let $o_i^t = \{g_j^{t-1}\}_{j \neq i}$ denote the outputs from the other agents from previous round, then the goal of agent i for this round is to generate an updated response $g_i^t \in \mathcal{G}$ that takes into account the responses from the other agents and it's own response from the previous round g_i^{t-1} , i.e.,

$$g_i^t \sim l_i(\mathcal{P}_{\text{debate}}(q, o_i^t, g_i^{t-1})),$$

where $\mathcal{P}_{\text{debate}}$ is the debate prompt.¹

This procedure runs over for T rounds, where the final output of the debate is the set of responses $\{g_i^T\}_{i=1}^n$. Then

¹If the responses are too large to fit in the context window, then they are summarized via making another LLM call, i.e., $o_i^t = \{l_i(\mathcal{P}_{\text{summarize}}(g_j^{t-1}))\}_{j \neq i}$.

the majority response across all agents is selected the final answer. Let $\text{majority} : \mathcal{G}^N \rightarrow \mathcal{A}$ denote the majority voting and filtering for responses across all agents, then the final answer is given by $a = \text{majority}(\{g_i^T\}_{i=1}^n)$.

4. Experimental Setup

We describe the set of tasks, models and prompts used in our experiments in this section.

4.1. Datasets

CommonSenseQA: The CommonSenseQA dataset (Tal- mor et al., 2019) consist of multiple-choice questions with complex semantics that often require prior knowledge to answer correctly. The dataset is intended to test for prior common-sense knowledge encoded within LLMs and check for common misconceptions.

MMLU: Massive Multitask Language Understanding, or MMLU (Hendrycks et al., 2021), is a widely used multiple-choice dataset covering 57 domains including elementary mathematics, US history, computer science, law, and more. To perform well on MMLU, models need robust world knowledge and problem solving ability.

GSM8K: GSM8K (Cobbe et al., 2021) is a dataset of linguistically diverse grade school math word problems which require multi-step mathematical reasoning to solve. This dataset is not multiple-choice and instead requires open-ended generation of the answer to the math questions, potentially with intermediate reasoning steps.

4.2. Models

We used three models from distinct model families in our experiments: GPT-4o-mini (OpenAI, 2024), LLaMA-3.1-8B-Instruct (et al., 2024) and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023). To align with prior work on multi-agent debate (Du et al., 2023; Subramaniam et al., 2025), we ran all experiments and models with default temperature parameter, the `top_p` = 0.9, maximum generation length of 2048 tokens and $T = 2$ rounds of debate. We take 100 random samples for each task from the dataset and report result over 5 random seeds.

4.3. Prompts

We provide examples of how the prompts for each task look for a random question below:

- **CommonSenseQA:** Can you answer the following question as accurately as possible? If a product doesn't last, what does it have a reputation of doing?: A) disintegrate, B) wear out, C) desolved, D) fall apart, E) dissipate Explain your answer by providing a bullet

point summary of your reasoning, putting the answer in the form (X) at the end of your response.

- **MMLU:** Can you answer the following question as accurately as possible? What is the value of p in $24 = 2p$? A) $p = 4$, B) $p = 8$, C) $p = 12$, D) $p = 24$ Explain your answer by providing a bullet point summary of your reasoning, putting the answer in the form (X) at the end of your response.
- **GSM8K:** Can you solve the following math problem? Mark is trying to choose between two venues for a surprise party for his wife. The first venue charges a flat fee of \$200, regardless of how many guests attend. While the second charges, \$25 per person who attends. However, the first venue does not include food, which Mark estimates will cost \$5 for each person who attends. At the second venue, food for each guest is already included in the price. How many guests are necessary for the two venues to be equal in cost? Provide a bullet point summary of your reasoning. Your final answer should be a single numerical number, in the form `[answer]`, at the end of your response.

For multi-agent debate, we additionally use the following prompt as $\mathcal{P}_{\text{debate}}$ for each round of debate, adjusted to the answer format of each task:

Debate prompt, $\mathcal{P}_{\text{debate}}$

These are the solutions to the problem from other agents: {AGENT_RESPONSES} Using the reasoning from other agents as additional advice, can you give an updated answer? Explain your reasoning. Examine your solution and that of other agents. Put your answer in the form (X) at the end of your response.

4.4. Code

We provide the source code for all our experiments at <https://github.com/TheNormativityLab/talk-aint-cheap/>.

5. Results

5.1. Effectiveness of Debate

We present results showing that debate can sometimes be *harmful* rather than helpful – in particular, that sometimes agents perform better *without any debate* than after exchanging reasons with other agents. We present the results in Table 1. We find that in case of CommonSenseQA, which was not studied in prior work on multi-agent debate (Du et al., 2023), debate almost always harms the performance. A key

insight is that even when groups include more “strong” models (e.g., GPT) than “weak” ones (e.g., Mistral), the process of debate does not always yield performance gains. Contrary to the prevailing narrative that debate improves collective reasoning, our experiments demonstrate that performance can actually decrease after agents engage in debate, even when stronger models outnumber weaker ones.

5.2. Performance Degradation during Debate

Figure 1 presents performance across three tasks—MMLU, CommonSenseQA, and GSM8K—as a function of debate rounds among groups of language models with varying individual performance on the underlying task. In fact, in many group configurations, we observe that performance *decreases* as the debate progresses. This trend appears across datasets, but is especially pronounced in MMLU and CommonSenseQA, where groups with mixed-capability models often suffer from group performance degradation during debate despite having a majority of stronger agents. These findings challenge the common assumption that deliberation or iterative reasoning among AI agents will always lead to better outcomes.

5.3. Does exchange of reasoning help in sequential revision?

If an LLM agent can reflect and correct mistakes based on the reasoning of other agents, we would expect the model to improve its answer and increase performance the collective performance of the group. We know the self-correction capability of single-agent LLMs does not easily work out of the box (Huang et al., 2023), and we want to evaluate if the debate procedure helps with this or not?

To assess this, we analyze how agent responses change between debate rounds across all our tasks. Note that there are four possible types of transitions: correct \rightarrow incorrect, correct \rightarrow correct, incorrect \rightarrow incorrect, and incorrect \rightarrow correct. As shown in Figure 2, we observe a significant shift in agent responses from correct to incorrect answers (shown in red), indicating that debate can actively mislead agents. Figure 3 highlights the trend that correct \rightarrow incorrect transitions occur more frequently than incorrect \rightarrow correct transitions (red $>$ green) across the runs. This suggests that stronger agents are more likely to flip from correct to incorrect answers compared to weaker agents learning from the reasoning and answers of stronger peers. Additionally, Figure 2 shows that the proportion of correct-to-incorrect transitions exceeds the proportion of incorrect-to-correct ones in subsequent rounds, corroborating our earlier results that debate performance degrades over rounds.

We hypothesize that this undesirable behavior may stem from RLHF (Kaufmann et al., 2024) post-training, which makes LLMs more sycophantic and compliant with user

























	CommonSense QA		MMLU		GSM8K	
1× 	74.8±1.9		82.6±2.1		93.2±1.8	
1× 	57.0±1.5		55.6±0.5		76.4±2.1	
1× 	41.6±2.3		34.0±1.9		34.2±1.8	
	w/o Debate	After Debate	w/o Debate	After Debate	w/o Debate	After Debate
3× 	44.4±2.7	39.4±3.9 ↓ 5.0	33.6±1.8	24.4±2.9 ↓ 9.2	43.6±1.5	46.4±1.4 ↑ 2.8
3× 	63.0±3.9	58.6±2.3 ↓ 4.4	61.6±2.4	57.8±1.8 ↓ 3.8	87.6±1.5	84.2±2.0 ↓ 3.4
3× 	75.6±2.2	74.8±2.1 ↓ 0.8	81.4±3.3	82.2±2.7 ↑ 0.8	94.0±0.9	94.4±1.5 ↑ 0.4
1×  , 2× 	66.2±2.2	64.4±2.1 ↓ 1.8	65.0±2.3	68.0±2.2 ↑ 3.0	88.4±1.3	92.8±1.7 ↑ 4.4
2×  , 1× 	74.8±1.3	74.0±0.7 ↓ 0.8	82.6±3.2	81.0±3.0 ↓ 1.6	93.6±0.8	94.6±1.3 ↑ 1.0
2×  , 1× 	58.2±3.8	50.2±3.9 ↓ 8.0	51.8±2.2	43.6±1.9 ↓ 8.2	82.6±1.9	75.8±2.1 ↓ 6.8
1×  , 2× 	53.4±2.7	46.8±2.5 ↓ 6.6	40.0±2.1	28.0±1.2 ↓ 12.0	61.0±2.4	64.8±1.5 ↑ 3.8
1×  , 2× 	62.4±1.1	59.4±1.9 ↓ 3.0	65.8±2.9	58.8±1.2 ↓ 7.0	90.2±0.8	87.8±1.9 ↓ 2.4
2×  , 1× 	74.6±1.6	72.4±2.7 ↓ 2.2	82.8±2.7	80.8±2.8 ↓ 2.0	93.4±1.3	93.0±1.3 ↓ 0.4
1×  , 1×  , 1× 	66.6±1.9	65.4±2.1 ↓ 1.2	57.8±3.1	63.4±2.1 ↑ 5.6	86.8±0.9	90.2±1.2 ↑ 3.4

Table 1: The first column shows the configuration of LLM agents, where  denotes GPT-4o-mini,  denotes Mistral-7B and  denotes the Llama-3.1-8B models. The top 3 rows benchmark the performance of how well a single agent performs on these tasks. The subsequent rows benchmark the effectiveness of debate procedure. The **w/o Debate** column represents the case where majority vote from multiple agents based on their initial responses is chosen, i.e., there is no exchange of reasoning. The **After Debate** denotes the result of majority vote after the exchange of reasonings via the debate procedure. The arrows denote the difference in performance post debate procedure characterizing the benefit of exchange of reasons on the performance. All the experiments were done on 100 random samples and across 5 different seeds, reported are mean and standard error.

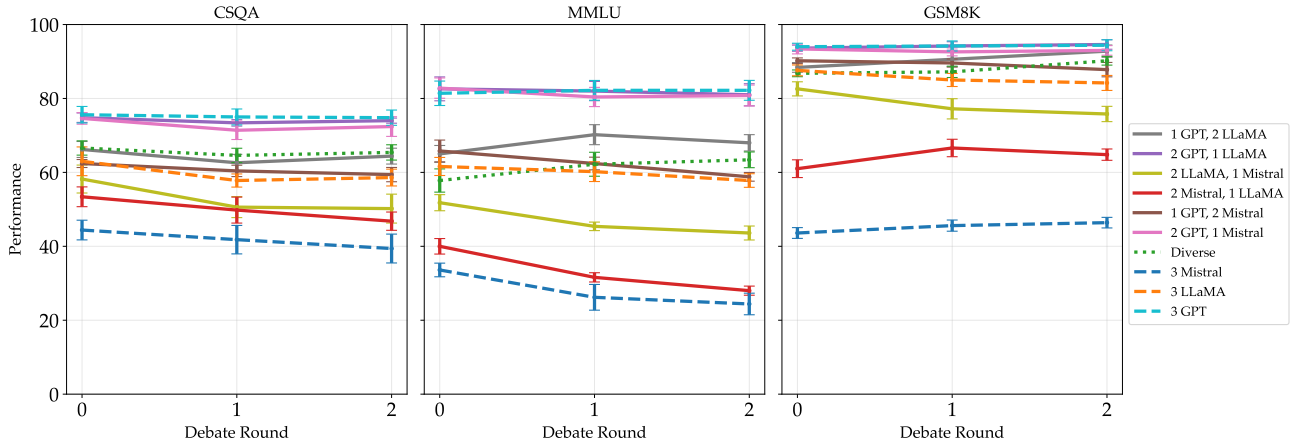





Figure 1: In many cases, we find that group accuracy frequently *degrades* over the course of debate, rather than improving performance. Diverse refers to the case (1× , 1× , 1× ).

opinions, leading them to adopt other agents’ answers, even when they may be incorrect. Sharma et al. (2023) demonstrate that humans and surrogate preference models sometimes prefer convincing but incorrect responses over terse correct ones, indicating that optimizing for human-like approval can sacrifice truthfulness in favor of sycophancy. In a debate setting, this suggests that agents might collude or

cater to each other rather than offer honest critique. If all agents are tuned to please a human (or each other), debates may devolve into polite agreement rather than productive critique.

Notably, sycophancy can cause strong models to yield to flawed arguments, resulting in a degradation of group performance during debates. We present some qualitative ex-

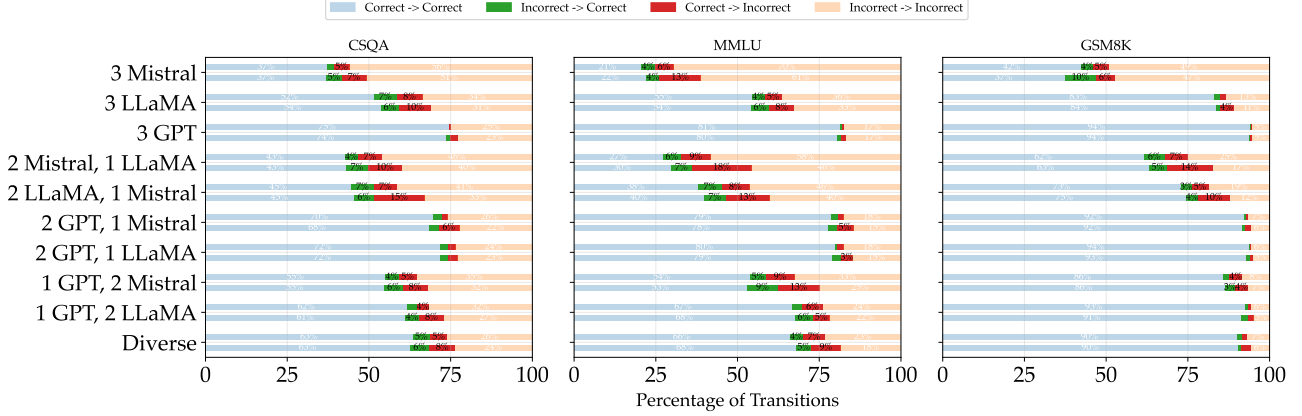


Figure 2: Breakdown of how agent change answers between debate rounds for different agent settings. The top row denotes the first round, and the row below denotes the second round. In almost cases we see red greater than green, i.e., we observe more correct-to-incorrect than incorrect-to-correct.

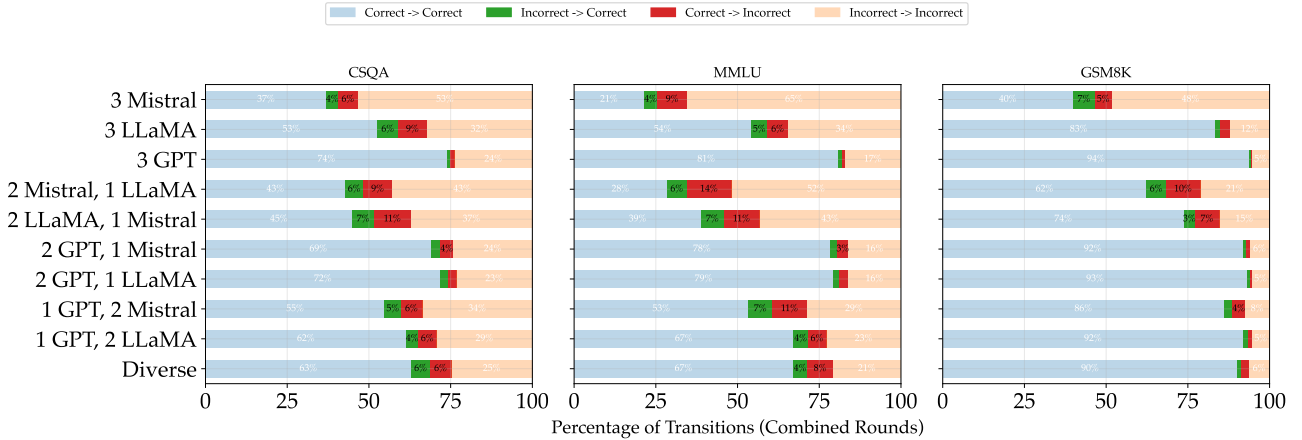


Figure 3: Breakdown of how agent change answers between debate rounds for different agent settings. The results are combined over rounds here.

amples where this behavior is evident in Appendix A.

Overall, these findings challenge the notion that debate inherently leads to better answers through shared reasoning. Instead, they suggest that without proper mechanisms to mitigate such degradations, debate can propagate errors even among otherwise capable agents.

6. Discussion

Our findings challenge the prevailing assumption that deliberation among AI agents will always improve reasoning. In fact, our experiments reveal that multi-agent debate can sometimes degrade performance: group accuracy often declines over successive rounds of debate. This counterintuitive trend holds even when the majority of agents perform well individually on the task. In other words, additional

exchange of reasons between agents does not always correct mistakes; instead, it can amplify them.

We frequently observe that stronger agents flip from correct to incorrect answers in response to weaker peers’ arguments more often than weaker agents learn the correct answer from their stronger peers. In these cases, stronger models agree reflexively with less capable models rather than critically evaluating their reasoning, and we confirm this observation by examining specific agent responses during debate. This behavior suggests that overly sycophantic behavior encouraged by current alignment techniques may inadvertently encourage undue deference, causing agents to cede their correct positions under pressure. The net effect is that heterogeneous groups can converge on wrong answers together, negating the potential benefits of discussion.

These results imply that naive debate protocols risk prop-

agating errors instead of correcting them. Our work highlights the need to design debate systems that discourage blind agreement through structured critique. For instance, future frameworks might actively encourage agents to consider the robustness of other agents' reasoning, or incorporate confidence estimates or credibility scores so that arguments are weighted by each agent's expertise. Training or incentive schemes could explicitly penalize unjustified agreement and reward independent verification of claims. By encouraging agents to selectively trust peer reasoning – rather than reflexively deferring – we could help preserve the constructive potential of multi-agent debate. Ultimately, improving debate robustness will require aligning models not just to agree politely, but to question peers rigorously when appropriate.

Acknowledgements

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners.

References

- Agarwal, M. and Khanna, D. When persuasion overrides truth in multi-agent llm debates: Introducing a confidence-weighted persuasion override rate (cw-por), 2025. URL <https://arxiv.org/abs/2504.00374>.
- Amayuelas, A., Yang, X., Antoniadis, A., Hua, W., Pan, L., and Wang, W. Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. *arXiv preprint arXiv:2406.14711*, 2024.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mor-datch, I. Improving factuality and reasoning in language models through multiagent debate, 2023. URL <https://arxiv.org/abs/2305.14325>.
- Estornell, A. and Liu, Y. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964, 2024.
- Estornell, A., Ton, J.-F., Yao, Y., and Liu, Y. Acc-collab: An actor-critic approach to multi-agent llm collaboration, 2025. URL <https://arxiv.org/abs/2411.00053>.
- et al., A. G. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Fu, Y., Peng, H., Khot, T., and Lapata, M. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.
- He, C., Zou, B., Li, X., Chen, J., Xing, J., and Ma, H. Enhancing llm reasoning with multi-path collaborative reactive and reflection agents, 2025. URL <https://arxiv.org/abs/2501.00430>.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.
- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.
- Irving, G., Christiano, P., and Amodei, D. Ai safety via debate, 2018. URL <https://arxiv.org/abs/1805.00899>.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- Kamoi, R., Zhang, Y., Zhang, N., Han, J., and Zhang, R. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.
- Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback, 2024. URL <https://arxiv.org/abs/2312.14925>.

- Kenton, Z., Siegel, N. Y., Kramár, J., Brown-Cohen, J., Albanie, S., Bulian, J., Agarwal, R., Lindner, D., Tang, Y., Goodman, N. D., and Shah, R. On scalable oversight with weak llms judging strong llms, 2024. URL <https://arxiv.org/abs/2407.04622>.
- Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers, 2024. URL <https://arxiv.org/abs/2402.06782>.
- Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. Camel: Communicative agents for "mind" exploration of large language model society, 2023. URL <https://arxiv.org/abs/2303.17760>.
- Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- Michael, J., Mahdi, S., Rein, D., Petty, J., Dirani, J., Padmakumar, V., and Bowman, S. R. Debate helps supervise unreliable experts, 2023. URL <https://arxiv.org/abs/2311.08702>.
- OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askeel, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models, 2023. URL <https://arxiv.org/abs/2310.13548>.
- Subramaniam, V., Du, Y., Tenenbaum, J. B., Torralba, A., Li, S., and Mordatch, I. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.
- Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL <https://arxiv.org/abs/1811.00937>.
- Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O’Sullivan, B., and Nguyen, H. D. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models, 2023. URL <https://arxiv.org/abs/2203.11171>.
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL <https://arxiv.org/abs/2308.08155>.
- Yang, H., Chen, J., Siew, M., Llorido-Botran, T., and Joe-Wong, C. Llm-powered decentralized generative agents with adaptive hierarchical knowledge graph for cooperative planning, 2025. URL <https://arxiv.org/abs/2502.05453>.

A. Example Sycophantic Model Responses

Here, we provide some specific instances in which we observe sycophantic behavior in models which change their answer from the correct answer to incorrect, in order to agree with another agent.

A.1. CommonSenseQA, 1 GPT and 2 LLaMA Agents

Example: CommonSenseQA, 1 GPT and 2 LLaMA Agents

Question:

What can a concert hall also be used for?: A) commercial building, B) university building, C) screaming, D) city, E) theater

Correct Answer: E

Initial Answer (GPT):

*After examining the reasoning from both agents and considering their perspectives, it's clear that concert halls can serve various purposes beyond hosting traditional music concerts...Based on this expanded analysis, the most appropriate option that captures an additional use of a concert hall is **(E) theater**, as it reflects the adaptability of these spaces to host both musical and theatrical performances, alongside other events. Thus, the final answer is (E).*

Modified Answer (GPT):

*After reviewing the insights from both agents and their reasoning regarding the uses of concert halls, I can provide an updated and more nuanced conclusion...Concert halls are primarily recognized for their role in hosting a wide variety of events, particularly theatrical performances but also encompassing other uses such as corporate events and educational activities. Therefore, while (E) theater is a strong candidate due to its specific relevance, I also acknowledge the multi-faceted nature of concert halls as emphasized by Agent 2. Thus, I conclude that the most fitting answer, reflecting the versatility and primary uses of concert halls, is **(B)** for their educational and corporate functionalities alongside their theatrical capacity, capturing the broader applicability discussed by Agent 2.*

Final answer: (B) ← Changed from correct to incorrect

A.2. GSM8K, 1 GPT, 1 LLaMA, 1 Mistral Agent

Example: GSM8K, 1 GPT, 1 LLaMA, 1 Mistral Agent

Question:

Jerry is rolling a six-sided die. How much more likely is it (expressed as a percentage) that he rolls a number greater than 3 than that he rolls two even numbers in a row?

Correct Answer: 25

Initial Answer (GPT):

First, we need to determine two probabilities: 1. The probability of rolling a number $n > 3$. 2. The probability of rolling two consecutive even numbers. There are 3 favorable outcomes (4, 5, 6) out of 6 possible outcomes. $P(n > 3) = 0.5$ and $P(\text{even}) = 0.5$. Then, $P(\text{two evens}) = P(\text{even}) \times P(\text{even}) = 0.25$ and $\text{Difference} = P(n > 3) - P(\text{two evens}) = 25\%$. Therefore, Jerry is 25% more likely to roll a number greater than 3 than to roll two evens in a row.

Modified Answer (GPT):

Let's solve the problem again using a clear approach and keeping in mind the insights provided by the other agents. $P(n > 3) = 3/6 = 0.5$, $P(\text{even}) = 3/6 = 1/2$, and $P(\text{even after even}) = 2/6 = 1/3$, so $P(\text{two evens}) = P(\text{even}) \times P(\text{even after even}) \approx 16.67\%$ and $\text{Difference} = P(n > 3) - P(\text{two evens}) \approx 33.33\%$. Therefore, Jerry is 33.33% more likely to roll a number greater than 3 than to roll two even numbers in a row.

Final answer: 33.33 ← Changed from correct to incorrect

A.3. MMLU, 2 LLaMA and 1 Mistral Agent

Example: MMLU, 2 LLaMA and 1 Mistral Agent

Question:

At the outset of group counseling, the therapist should explain that the obligation of the group member to respect confidentiality is: A) less of an issue than it is in individual therapy, B) the legal responsibility of the therapist, C) protected by state/provincial laws, D) on the "honor system" only

Correct Answer: D

Initial Answer (LLaMA):

...Ultimately, maintaining confidentiality in group counseling requires a collaborative approach, where both the therapist and group members take responsibility for respecting confidentiality. **D) on the "honor system" only**

Modified Answer (LLaMA):

After examining the opinions of all other agents, I've come to a revised conclusion...A collective commitment to respect confidentiality is crucial in group counseling, and both the therapist and group members must take responsibility for maintaining confidentiality. **B) the legal responsibility of the therapist**

Final answer: (B) ← Changed from correct to incorrect