Improving Multi-Agent Debate with Sparse Communication Topology

Yunxuan Li[1†], Yibing Du[1], Jiageng Zhang[1], Le Hou[2],
Peter Grabowski[1], Yeqing Li[1], Eugene Ie[1]
[1]Google [2]Google DeepMind

Abstract

Multi-agent debate has proven effective in improving large language models quality for reasoning and factuality tasks. While various role-playing strategies in multi-agent debates have been explored, in terms of the communication among agents, existing approaches adopt a brute-force algorithm - each agent can communicate with all other agents. In this paper, we systematically investigate the effect of communication connectivity in multi-agent systems. Our experiments on GPT and Mistral models reveal that multi-agent debates leveraging sparse communication topology can achieve comparable or superior performance while significantly reducing computational costs. Furthermore, we extend the multi-agent debate framework to multimodal reasoning and alignment labeling tasks, showcasing its broad applicability and effectiveness. Our findings underscore the importance of communication connectivity on enhancing the efficiency and effectiveness of the "society of minds" approach.

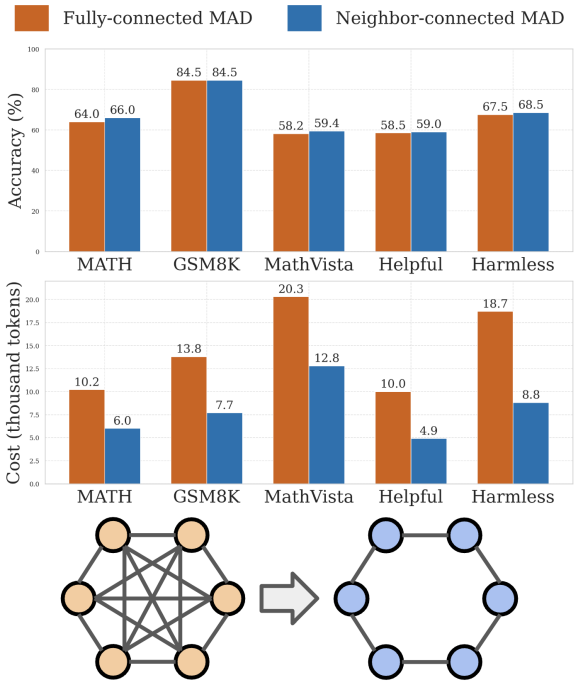

Figure 1: Accuracy (top) and inference input cost (middle) comparison of multi-agent debate system between fully-connected (bottom left) and neighbor-connected (bottom right) communication topologies.

1 Introduction

Large language models (LLMs) have demonstrated exceptional performance in natural language understanding and generation tasks. Recently a paradigm shift towards prompting LLMs has emerged as a significant and influential research area. By leveraging the in-context learning (ICL) capabilities of LLMs, these models can be adapted to various tasks such as reasoning, factuality, and AI feedback.

Several prompting methods have been developed to enhance LLM performance by optimizing their ICL capabilities. Notable techniques include Chain-of-Thought (CoT) (Wei et al., 2022), self-consistency (SC) (Wang et al., 2022), and self-critique (Madaan et al., 2024; Welleck et al., 2022; Shinn et al., 2024). Recently, the multi-agent debate (MAD) framework is proven to be an innovative approach. Similar to a human discussion process, MAD employs multiple LLM agents to engage in discussions with one another, combining their reasoning and critical thinking abilities to produce high-quality results. Specifically, given a question, each agent first generates their own solutions to the question and then takes other agents' solutions as reference to update its own answer. This process can be repeated for several rounds. MAD has demonstrated significant improvement on factuality and reasoning tasks. While the debate process is highly productive, it is also very costly: As the number of LLM agents and debate rounds increase, the input context expands significantly.

[†]Correspondence: yunxuanli@google.com

Findings of the Association for Computational Linguistics: EACL 2024, pages 7281–7294

Inspired by the intensive computational cost of MAD, a natural question arises: *What if we reduce the number of reference solutions visible to each agent?* We conduct a systematic study on the sparsity of the multi-agent communication topology. Surprisingly, we find that sparse communication connectivity can deliver comparable or superior performance while significantly reducing inference costs. Figure 1 presents a comparison between fully-connected MAD and neighbor-connected MAD. Compared to fully-connected MAD, neighbor-connected MAD achieves an improvement of $+2\%$ on the MATH dataset and maintains the same accuracy on GSM8K. Meanwhile, the average input token cost for reasoning tasks is reduced by over 40%.

MAD can also be a promising approach for Reinforcement Learning with AI Feedback (RLAIF) (Bai et al., 2022b; Lee et al., 2023) and weak-to-strong generalization (Burns et al., 2023). By delivering better reward signals, MAD has the potential to significantly aid in aligning large language models. To assess this, we first extend the MAD framework to alignment labeling tasks, demonstrating its effectiveness compared to single-agent setups. Additionally, we verify that the advantages of sparsity observed in the reasoning tasks experiments also apply to alignment labeling tasks. Our experiments on the Anthropic-HH datasets show an improvement of $+0.5\%$ in helpfulness and $+1.0\%$ in harmlessness, while reducing costs by 50.0% and 53.3%, respectively.

We find that when agents are instantiated by different LLMs within the MAD framework, interactions among multiple LLMs allow weaker models to be progressively strengthened through engagement with stronger models. In non-regular graph settings, assigning stronger LLMs to agents with higher centrality consistently yields better performance.

In summary, our contributions are listed as follows: (1) We demonstrate that sparse communication topology enhances both effectiveness and efficiency of the multi-agent debate framework; (2) We thoroughly evaluate sparse MAD for text-only and multimodal reasoning tasks, showing its advantage over standard MAD; (3) We extend the MAD framework to alignment labeling tasks, showing the effectiveness of standard MAD and further performance improvement introduced by sparse MAD; (4) We provide insights that explain the effectiveness of sparsity in MAD; (5) We find that assigning stronger LLMs to agents with higher centrality yields better overall performance in multiple LLM debate setup.

2 Related Work

Multi-Agent Debate MAD utilizes multiple LLM agents to discuss and debate with each other to generate and update the responses. It was first introduced by Du et al. (2023). Most of the MAD work focus on diversifying agents during the debate process. Liang et al. (2023); Park et al. (2023); Li et al. (2023a); Chan et al. (2023) highlight the importance of assigning different roles for agents. Chen et al. (2023) diversifies agents' responses by instantiated with multiple LLMs. Wang et al. (2024a) proposes a method in which agents are divided into sub-groups and their discussion outcomes are later merged. Qian et al. (2023); Wu et al. (2024); Hong et al. (2024) demonstrate the advantage of multi-agent collaboration in solving complex tasks. Unlike other work, we aim to explore the effectiveness of sparse communication topology in MAD, and extend its applications to reasoning and alignment tasks.

LLM Reasoning Much work has been done to improve the reasoning ability of language models with prompting, including Chain-of-Thought (Wei et al., 2022) and its variants (Yao et al., 2024; Besta et al., 2024), problem decomposition (Zhou et al., 2022), reasoning ensemble (Wang et al., 2022), reasoner with verification (Cobbe et al., 2021; Wang et al., 2024b; Luo et al., 2023).

Multimodal Reasoning With the recent advancements in vision-language models (Radford et al., 2021; Yu et al., 2022; Li et al., 2023b; Liu et al., 2024; Lin et al., 2024), multimodal large-language models (MLLMs) have demonstrated exceptional visual understanding capabilities. Several evaluation benchmarks have been proposed, such as VQAv2 (Goyal et al., 2017), OK-VQA (Marino et al., 2019), ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2023), and MathVista (Lu et al., 2023). Similar to LLMs, MLLMs can also be improved through prompt-based methods. Various attempts have been made to enhance MLLMs in this manner (Zheng et al., 2024; Ganz et al., 2024; Yang et al., 2023; Zhao et al., 2024; Zhang et al., 2023; Chen et al., 2024; Hu et al., 2024). Despite the effectiveness of these methods, they are often complex to design and implement. In this paper, we focus on improving multimodal reasoning using a

multi-agent approach.

AI Feedback Bai et al. (2022b) first introduces the idea of RLAIF, in which LLM is used to annotate harmlessness preference. Lee et al. (2023) compares various AI annotation methods. Recent work (Guo et al., 2024) also explores using AI feedback for online reinforcement learning, demonstrating the advantage of AI feedback for alignment research.

3 Method

3.1 Communication Topology

Communication topology of MAD refers to the connectivity structure among agents during the debate process. Communication topology can be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of agents and $\mathcal{E}$ is a set of communication channel. Presence of of any (e_i, e_j) in $\mathcal{E}$ indicates that agent i can access agent j's previous round solutions during the debate process, and vice versa. We focus on static graphs in this work, while we also did exploratory experiments with dynamic graphs (Appendix E).

We quantify the density of these graphs using the ratio of the number of edges to the maximum possible number of edges

$$D = \frac{2|\mathcal{E}|}{|\mathcal{V}|(|\mathcal{V}| - 1)}$$

A lower value of D indicates a sparser graph. In the standard MAD framework, agents are fully connected with each other, resulting in $D = 1$. In contrast, a neighbor-connected MAD has $|\mathcal{E}| = |\mathcal{V}|$, yielding $D = \frac{2}{|\mathcal{V}|-1}$, which is a sparse graph. While the findings of this paper can be generalized to communication topology with an arbitrary number of agents, we focus on regular graphs where all agents have same degrees and are permutation invariant, with $|\mathcal{V}| = 6$ (Figure 2). This choice is due to the limited spectrum of sparsity in scenarios with fewer agents and the significantly higher computational costs associated with analyzing scenarios with more agents. Additional experiment results with $|\mathcal{V}| = 4$ is shown in Appendix D.

3.2 Multi-Agent Debate Process

A typical MAD framework includes three steps:

(1) Individual Response Generation: In round 1, agents are initialized with LLMs, and then independently generate solutions to a given question.

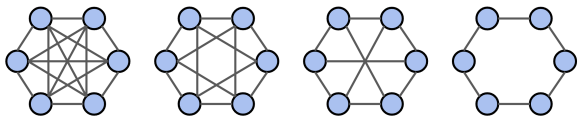

Figure 2: Communication topology of 6 agents with various sparsity. From left to right, the densities are 1 (fully-connected), $\frac{4}{5}$, $\frac{3}{5}$, and $\frac{2}{5}$ (neighbor-connected).

Typically a random decoding strategy is applied to diversify the solutions generated by agents.

(2) Multi-agent Debate: Starting round 2, each agent incorporates the responses of its connected peers from the previous round to critique or refine its own response. We utilize the standard *Simultaneous-Talk* communication strategy (Chan et al., 2023) to facilitate asynchronous computation. This debating process can occur over multiple rounds.

(3) Reaching Consensus: After the debate process, agents may still have differing solutions. In such cases, a majority vote is conducted among all agents to determine a consensus solution.

4 Experiments Setup

4.1 Tasks

We aim to validate the effectiveness and efficiency of sparse MAD on reasoning and alignment labeling tasks. For reasoning tasks, we consider two text-only reasoning tasks and one multimodal reasoning task: (1) MATH (Hendrycks et al., 2021): an arithmetic reasoning task containing challenging competition mathematics problems. We only use the *algebra linear 1d composed* sub-task for simplicity. (2) GSM8K (Cobbe et al., 2021): a high quality grade school math reasoning task. (3) MathVista (Lu et al., 2023): a benchmark designed to combine challenges from diverse mathematical and visual tasks. We only choose from *free_form* question type for consistency. For alignment labeling tasks, we consider Anthropic-HH dataset (Bai et al., 2022a): human preference data on helpfulness and harmlessness.

4.2 Models

Our experiments utilize three publicly available models: GPT-3.5 (OpenAI, 2022), GPT-4 (OpenAI, 2023), and Mistral 7B (Jiang et al., 2023). Specifically, we employ GPT-3.5 for text-only reasoning tasks and GPT-4 for multimodal reasoning tasks. For alignment labeling tasks, we use both GPT-3.5 and Mistral 7B. We refrain from using

GPT-4 for other tasks due to its significantly higher cost, which is approximately 10 times that of GPT-3.5. Additionally, we do not employ Mistral 7B for other tasks because of its inferior zero-shot performance on arithmetic reasoning. We randomly select 100 examples for each experiments involving GPT, and 500 examples for experiments with Mistral 7B.

4.3 Baselines

Our MAD setup employs 6 agents and engages them in debate for 5 rounds. We compare sparse MAD against the following baselines:

(1) Chain-of-Thought (**CoT**): CoT prompting improves reasoning capabilities of LLMs with explicit intermediate reasoning steps.

(2) Self-consistency (**SC**): SC margins out intermediate reasoning paths by sampling diverse reasoning paths and selecting the most consistent answer. We sampled 6 responses for SC, where each agent generates one response and we determine the final output by majority voting.

(3) Existing MAD (**MAD** ($D = 1$)): the standard approach for multi-agent debate, in which agents can communicate with all other agents with simultaneous-talk strategy. We also denote it as fully-connected MAD.

4.4 Evaluation Metrics

For reasoning tasks, we use the accuracy with respect to the ground truth answer to measure the quality of MAD. For alignment labeling tasks, we use *AI Labeler Alignment* (Lee et al., 2023) to measure the accuracy of MAD labeling with respect to the human annotation.

Cost refers to the input inference cost of LLMs, which typically involves handling the autoregressive decoding mechanism and computational resources. Considering that advanced LLMs use a pay-per-token pricing model, we measure the inference cost by the number of input tokens. Notably, while the input token cost is influenced by the topology design, the output token cost remains unaffected by it. As a result, we focus exclusively on reporting input cost savings in our results.

4.5 Variance Reduction

Evaluating the significance of new communication topology compared to existing one typically involves running multiple random experiments to estimate the mean and variance of performance. However, this approach becomes impractical when the signal-to-noise ratio is low and each experimental run is computationally expensive. To address this, we employ two methods to reduce experimental variance and enhance the sensitivity of MAD with respect to the changes in communication topology: (1) As used by Wang et al. (2024a), we reduce the temperature during language model decoding to stabilize performance. While we use the default temperature settings in API calls for most tasks, we lower the temperature to 0.25 for text arithmetic reasoning tasks to ensure robustness. (2) We employ conditional variance reduction (Ross, 2002). Observing that most of the variance arises from the first round of individual responses, we first generate a set of initial agent responses and then fix them in all subsequent debate processes across various communication topology designs. This approach effectively minimizes variance and improves the reliability of our experimental results.

5 Experiments: MAD with Single LLM

5.1 MAD on Text Reasoning Tasks

We build on existing work on MAD, exemplified by reasoning tasks, by showing the advantages of sparse MAD on top of the proven advantage of fully-connected MAD. Sparse MAD significantly saves computational cost while preserving comparable or better performance.

Sparse MAD has similar or higher accuracy with significant cost saving on reasoning tasks: For both the MATH and GSM8K tasks, we demonstrate that sparse MAD produces comparable or better accuracy than fully-connected MAD, while significantly cutting down inference costs. Both fully-connected and sparse MAD setups outperform Chain-of-Thought and self-consistency methods. Specifically, in the MATH task, fully-connected MAD shows a $+4.0\%$ quality gain over self-consistency, while sparse MAD configurations achieve accuracy improvements ranging from $+3.0\%$ to $+7.5\%$ (Table 1). Similarly, in the GSM8K task, fully-connected MAD demonstrates a $+4.5\%$ quality gain over self-consistency, whereas sparse MAD achieves accuracy improvements between $+3.5\%$ and $+6.5\%$ (Table 2). Furthermore, sparse MAD setups reduce costs by up to -41.5% and -43.5%, respectively. It is important to note that we exclusively use the GPT-3.5 model because Mistral 7B performs poorly on these challenging tasks in a zero-shot setting. More experiments on text-reasoning tasks are shown in

Method	Accuracy	Cost Saving
CoT	58.0 ± 2.0	-
SC	60.0	-
MAD ($D = 1$)	64.0 ± 1.4	baseline
MAD ($D = 4/5$)	$\mathbf{67.5 \pm 2.0}$	-14.6%
MAD ($D = 3/5$)	63.0 ± 1.8	-29.2%
MAD ($D = 2/5$)	66.0 ± 2.3	-41.5%

Table 1: Comparison of accuracy and cost savings of MAD against baseline methods on the MATH dataset. All experiments were conducted using the GPT-3.5 model.

Method	Accuracy	Cost Saving
CoT	77.5 ± 4.2	-
SC	80.0	-
MAD ($D = 1$)	84.5 ± 1.5	baseline
MAD ($D = 4/5$)	83.5 ± 0.5	-12.7%
MAD ($D = 3/5$)	$\mathbf{86.5 \pm 1.5}$	-29.1%
MAD ($D = 2/5$)	84.5 ± 0.8	-43.6%

Table 2: Comparison of accuracy and cost savings of MAD against baseline methods on the GSM8K dataset. All experiments were conducted using the GPT-3.5 model.

Appendix B.

5.2 MAD on Multimodal Reasoning Task

MAD on multimodal reasoning tasks also demonstrates notable improvements compared to Chain-of-Thought and self-consistency approaches. This suggests that MLLMs like GPT-4o can effectively integrate step-by-step reasoning with visual content to enhance final answers. Similar to text reasoning experiments, we examine various sparse MAD configurations and report their performance.

Sparse MAD retains performance while introducing significant cost savings on multimodal reasoning tasks. For the MathVista task, we evaluate different MAD configurations, comparing them to each other as well as to Chain-of-Thought (CoT) and self-consistency methods (Table 3). We find that sparse MAD achieves similar or slightly better accuracy compared to fully-connected MAD, with both outperforming CoT and self-consistency. The best sparse MAD configuration achieves a $+1.2\%$ improvement over fully-connected MAD and a $+6.4\%$ improvement over self-consistency. Additionally, sparse MAD provides substantial cost savings, reducing the total number of tokens used by up to 33.1%. Given that multimodal inputs are typically much larger than textual inputs (e.g., in GPT-4o, each image costs at least 225 tokens and can grow to 400+, 600+, or more tokens), we observe a total reduction of 40.6% in token usage, excluding the input image tokens.

Method	Accuracy	Cost Saving
CoT	52.4 ± 2.6	-
SC	53.0	-
MAD ($D = 1$)	58.2 ± 1.5	baseline
MAD ($D = 4/5$)	57.8 ± 1.9	-9.1% (-11.5%)
MAD ($D = 3/5$)	55.4 ± 0.9	-20.0% (-24.7%)
MAD ($D = 2/5$)	$\mathbf{59.4 \pm 0.6}$	-33.1% (-40.6%)

Table 3: Comparison of accuracy and cost savings of MAD against baseline methods on the MathVista dataset. All experiments were conducted using the GPT-4o model with the default temperature $T = 1$. The cost saving percentages in parenthesis are computed without multimodal inputs.

5.3 MAD on Alignment Labeling Tasks

Alignment labeling tasks involve annotating preferences between pairs of responses generated for a given question. Our prompt consists of three parts: (1) a system prompt that informs the LLM of its role as a rater and specifies the required answer formatting; (2) a question description providing the context of the question; and (3) an ending instruction that constrains the answer length and reiterates the formatting requirements. During the debate, reference solutions are included before the ending instruction. See A for more details.

We use *AI Labeler Alignment* (Lee et al., 2023) to measure the accuracy of MAD labeling with respect to the human annotation. To prevent potential position bias, we randomly assign the chosen response to either the (A) or (B) option. We report the accuracy and inference cost of MAD with various level of sparsity in Table 4 for helpfulness and Table 5 for harmlessness.

MAD outperforms single-agent on alignment labeling tasks: We find that MAD consistently outperforms single-agent methods, including CoT and self-consistency. On the helpfulness task, fully-connected MAD achieves a $+1.5\%$ and $+2.9\%$ improvement over self-consistency for GPT-3.5 and Mistral 7B models, respectively. On the harmlessness task, fully-connected MAD achieves a $+0.5\%$

Method	GPT-3.5		Mistral 7B	
	Accuracy	Cost Saving	Accuracy	Cost Saving
CoT	56.5 ± 3.1	-	60.8 ± 1.2	-
Self-Consistency	57.0	-	62.6	-
MAD ($D = 1$)	58.5 ± 1.7	baseline	65.5 ± 0.6	baseline
MAD ($D = 4/5$)	$\mathbf{59.0 \pm 1.8}$	-17.5%	65.6 ± 0.9	-18.3%
MAD ($D = 3/5$)	57.0 ± 1.6	-32.5%	64.6 ± 0.6	-35.2%
MAD ($D = 2/5$)	$\mathbf{59.0 \pm 1.4}$	-50.0%	$\mathbf{66.6 \pm 0.5}$	-53.5%

Table 4: AI labeler alignment accuracy and cost savings of MAD compared with baselines on the helpfulness dataset for GPT-3.5 and Mistral 7B models.

Method	GPT-3.5		Mistral 7B	
	Accuracy	Cost Saving	Accuracy	Cost Saving
CoT	66.0 ± 4.8	-	58.2 ± 2.0	-
Self-Consistency	67.0	-	60.0	-
MAD ($D = 1$)	67.5 ± 0.6	baseline	60.7 ± 0.3	baseline
MAD ($D = 4/5$)	67.0 ± 0.8	-17.3%	$\mathbf{62.2 \pm 0.2}$	-17.9%
MAD ($D = 3/5$)	67.5 ± 1.0	-34.7%	60.4 ± 0.4	-34.3%
MAD ($D = 2/5$)	$\mathbf{68.5 \pm 0.7}$	-53.3%	61.7 ± 0.2	-52.2%

Table 5: AI labeler alignment accuracy and cost savings of MAD compared with baselines on the harmlessness dataset for GPT-3.5 and Mistral 7B models.

and $+0.7\%$ improvement over self-consistency for GPT-3.5 and Mistral 7B models, respectively. These results suggest that the additional debate process in MAD, followed by majority voting, allows agents to incorporate perspectives from others and refine their opinions toward the correct answers during the debate process.

Sparse MAD can perform better with lower inference costs: Most sparse MAD configurations perform as well as or better than the fully-connected MAD, with at least one sparse topology outperforming the fully-connected MAD. Depending on the task, sparse MAD with GPT-3.5 can enhance performance by approximately $+0.5\%$ to $+1.0\%$, and sparse MAD with Mistral 7B can improve performance by about $+1.1\%$ to $+1.5\%$. Additionally, sparse MAD can reduce costs by up to -53.3% and -53.5%, respectively.

We observed that GPT-3.5 exhibits lower alignment accuracy compared to Mistral 7B on the helpfulness task. We attribute this discrepancy to the differences in pre-training and post-training corpora between the two models, which may lead to varying default preferences in a zero-shot setting. While we hypothesize that few-shot prompting techniques could mitigate this issue, exploring this is beyond the scope of this work.

5.4 Why Does Sparse MAD Work?

The common explanation for the effectiveness of MAD against single-agent setups is that agents can consider different perspectives before arriving at an answer. However, our experiment on the effectiveness of sparse MAD seems challenge this intuition. In this section, we aim to explain why sparse MAD can achieve comparable or even superior performance.

Impact of incorrect reference solutions: In a MAD framework, we define $Q(n, p)$ as the probability that a single agent delivers correct answers given n reference solutions, where p percentage of these are correct. This probability, $Q(n, p)$, can be estimated using Monte Carlo sampling with constructed in-context reference solutions. As a case study, we selected three questions from the GSM8K dataset and estimated $Q(n, p)$ for $n \in \{2, 3, 4, 5\}$ and $p \in \{0\%, 25\%, 50\%, 75\%, 100\%\}$. Here, the choice of n corresponds to the single-agent scenarios in MAD with $D = \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1$. Results shown in Figure 3 indicate that for easier questions, where most reference solutions are correct, an increase in the number of observed reference solutions (namely MAD becomes denser) improves the likelihood of the agent arriving at the correct answer. Conversely, for more difficult questions, where most agents do not provide correct answers,

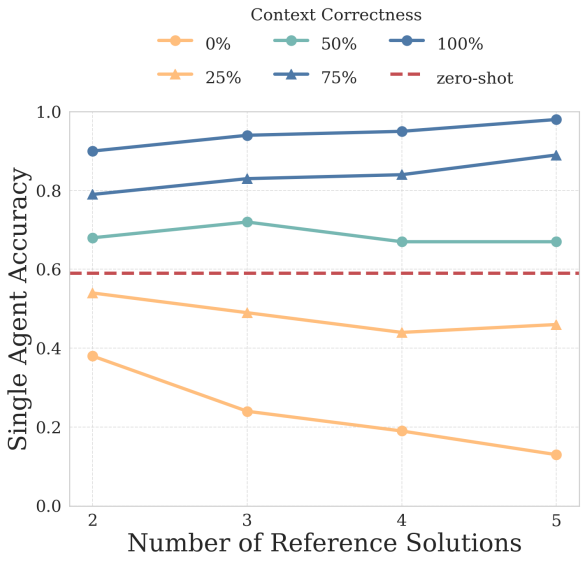

Figure 3: Probability of a single agent generating correct answers given n reference solutions, with p representing the correctness of these solutions. Monte Carlo sampling was performed on three questions, each with 100 runs.

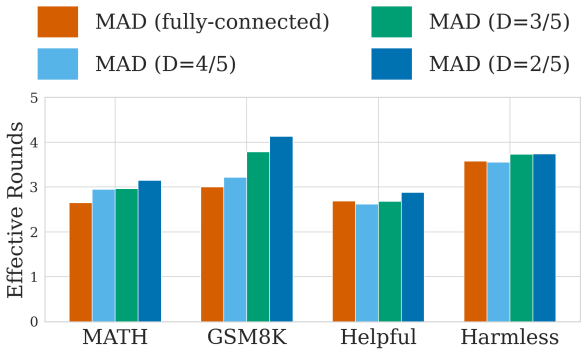

Figure 4: Effective debate rounds for each topology design in reasoning and alignment labeling tasks.

an increase in the number of observed reference solutions tends to mislead the agent into choosing incorrect answers, thereby drastically reducing the likelihood of reaching a correct response.

Sparser MAD allows more rounds of effective debate: We observe that once all agents converge on the same answer, it becomes highly unlikely for any of them to change their decision. Sparser MAD addresses the convergence issue, a primary limitation observed in fully-connected MAD. We define the number of effective debates as the number of rounds before all agents reach the same answer. Figure 4 illustrates the effective number of debate rounds for various topologies in reasoning and alignment labeling tasks. Our results show that sparse MAD tends to sustain longer debates before achieving consensus, indicating that sparse MAD allows for more extensive deliberation and in-depth discussion by preventing premature convergence and encouraging a broader exploration of potential solutions. We observe there are similar findings in the Chain-of-Thought prompting (Jin et al., 2024) and MAD (Du et al., 2023) that the increase of reasoning length can significantly improve the performance.

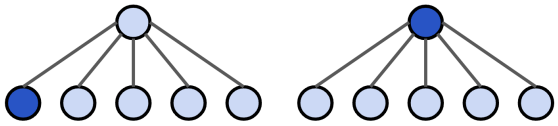

Figure 5: Isotropic communication topology with two setups: the stronger LLM has low centrality (left) and high centrality (right).

Centrality	Accuracy	
	SC	Isotropic MAD
High	64.0	**67.0 ± 0.8**
Low	64.0	65.8 ± 0.5

Table 6: Comparison of accuracy depending on where a stronger LLM is placed in debate, using the Harmlessness task as example. In both cases, there are 5 Mistral models and 1 GPT-3.5 Model. Accuracy of Isotropic MAD is calculated as the average over debate rounds.

6 Experiments: MAD with Multiple LLMs

Previous sections focus on the MAD with agents instantiated by the same LLM. In this section, we explore the scenario when multiple LLMs are available. With agents instantiated by different LLMs, the permutation invariance symmetry is broken, and the regular graph may not be optimal. A natural question is: *how to design the communication topology given a MAD framework of N agents, in which M instantiated by the stronger LLM and $N - M$ instantiated by the weaker LLM?*

Assigning stronger LLMs to agents with higher centrality yields better performance: We conducted experiments on harmlessness alignment labeling task, involving 6 agents, with 1 agent utilizing GPT-3.5 (the stronger LLM) and the remaining 5 agents utilizing Mistral 7B (the weaker LLM). We tested two setups on the isotropic communication topology: one where the stronger LLM had a degree of 1 (indicating low centrality) and another where it had a degree of 5 (indicating high centrality), as illustrated in Figure 5. The experimental

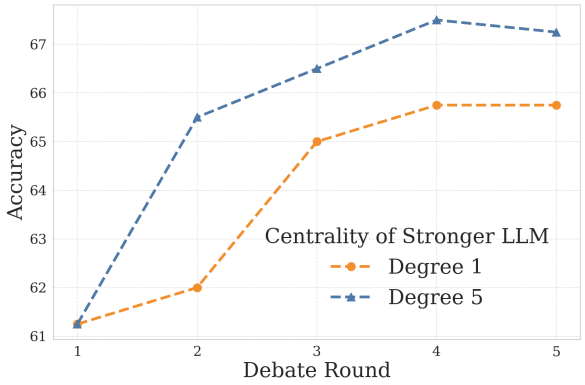

Figure 6: Average accuracy of weaker agents across different debate rounds.

results presented in Table 6 show that positioning the stronger LLM at a node with higher centrality (degree of 5) leads to better performance ($+3.0\%$ improvement) compared to placing it at a node with lower centrality (degree of 1) which resulted in a $+1.8\%$ improvement.

The results above underscore the importance of information flow in the design of communication topology. Figure 6 illustrates the average accuracy of weaker agents with respect to the number of debate rounds. When the stronger agent has a degree of 5, it can effectively disseminate its knowledge to weaker agents in just one debate round, resulting in a sharp increase in the average accuracy of weaker LLMs. In contrast, when the stronger agent has a degree of 1, the process requires two rounds: first, the information is transmitted to the central weaker agent in the first debate round (round 2), which then shares it with other weaker agents in the next round (round 3). This two-step process leads to greater information loss.

7 Conclusion

In this paper, we show that sparse communication topologies can improve the multi-agent debate framework significantly. Our results indicate that sparse MAD configurations achieve comparable or superior performance to standard MADs while greatly reducing computational costs. We also extend the MAD framework to alignment labeling tasks, demonstrating the benefits of MADs over single-agent setups and self-consistency and further highlighting the benefits of sparse MADs over fully-connected configurations. We present case-study insights that explain the effectiveness of sparse MADs. Additionally, we investigate the impact of communication topology design with multiple large language models (LLMs), finding that assigning stronger LLMs to more connected agents enhances overall performance.

In summary, our work paves the way for more efficient and effective multi-agent systems by leveraging sparse communication topologies. Future studies could focus on deepening our understanding of the underlying mechanisms and developing strategies for optimal topology design in multi-agent frameworks.

8 Ethical Considerations

In this work, several ethical considerations were addressed to ensure the integrity and responsible use of the system:

Public Datasets: The framework was built using publicly available datasets that are designed for academic research. We strictly adhered to ethical guidelines by not using any personal or confidential data.

License: Only public APIs that offer appropriate licensing were utilized. This ensures that all external tools are used in a lawful and ethical manner.

AI assistant: AI tools were employed solely for polishing writing and correcting grammar. The AI was not used to generate content or ideas, maintaining the authenticity and originality of the research work.

9 Limitations

While our study provides valuable insights into the communication topology analysis of multi-agent debate, several limitations must be acknowledged:

Our analysis is primarily based on static graphs where the communication topology remains unchanged throughout the debate rounds. This constraint simplifies the analysis, but ignores the dynamic nature of real-world communication networks. Additionally, our study focuses on prompt design under a zero-shot setting, utilizing only publicly available GPT and Mistral models. This narrow scope may not fully capture the variability and adaptability present in more diverse agent populations. Furthermore, we confined our analysis to regular graphs, which do not encompass the full spectrum of potential graph configurations. Future work should consider dynamic graphs, a broader range of models, and varied graph connectivity to better reflect the evolving and complex nature of multi-agent interactions.

Our study relies on a subset of academic datasets due to limited data access as well as computational constraints. While these datasets provide a valuable foundation for analyzing communication graph dynamics in multi-agent debates, they may not fully represent the diversity and complexity found in broader real-world data. The restricted scope limits our ability to generalize findings across different domains and contexts. Future research should aim to include a wider range of datasets, potentially leveraging more efficient computational resources, to enhance the robustness and applicability of our findings.

We lack a rigorous theoretical proof explaining why sparse connectivity can lead to better performance. This gap in our understanding limits our ability to generalize our findings and apply them with confidence in various settings. Secondly, we do not have a definitive method for determining the optimal topology design, which is crucial for maximizing the efficiency and effectiveness of multi-agent systems. Addressing these questions is essential for future research. Potential explanations might involve theoretical insights, social and psychological dynamics, or a combination of these factors. Additionally, fine-tuning models could offer further clarity and aid in optimizing communication topology. Future work should aim to develop robust theoretical frameworks and empirical strategies to better understand and leverage communication topology in multi-agent debates.

The multi-agent debate framework holds significant potential for various real-world applications. However, it also carries the risk of misuse, including the dissemination of biased information or misinformation. Additionally, the framework requires substantial computational resources, which could impact energy consumption and environmental sustainability. Future research should focus on developing robust, trustworthy, and energy-efficient multi-agent systems to mitigate these risks and ensure ethical, reliable, and sustainable outcomes.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2024. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems, 2021. *URL https://arxiv.org/abs/2110.14168*.

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

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13861–13871.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Liangchen Luo, Zi Lin, Yinxiao Liu, Lei Shu, Yun Zhu, Jingbo Shang, and Lei Meng. 2023. Critique ability of large language models. *arXiv preprint arXiv:2310.04815*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt 4 technical report. *arXiv preprint arXiv:2303.08774*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

S.M. Ross. 2002. *Simulation*. Academic Press.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024a. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. 2024b. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. *arXiv preprint arXiv:2402.02658*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *COLM*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Xueliang Zhao, Xinting Huang, Tingchen Fu, Qintong Li, Shansan Gong, Lemao Liu, Wei Bi, and Lingpeng Kong. 2024. Bba: Bi-modal behavioral alignment for reasoning with large vision-language models. *arXiv preprint arXiv:2402.13577*.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. 2024. A picture is worth a graph: Blueprint debate on graph for multimodal reasoning. *arXiv preprint arXiv:2403.14972*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Prompt Templates

A.1 Text Reasoning Tasks

System Prompt:
You are a helpful assistant with expertise in mathematics and reasoning. Your task is to assist in solving a math reasoning problem by providing a clear and detailed solution. Limit your output within 100 words, and your final answer should be a single numerical number, in the form of {{*answer*}}, at the end of your response.
Starting Prompt:
Can you solve the following math problem? {*question*} Explain your reasoning. Your final answer should be a single numerical number, in the form of {{*answer*}}, at the end of your response.
Debate Prompt:
These are the solutions to the problem from other agents:
One agent solution: {*reference solution*}
One agent solution: {*reference solution*}
One agent solution: {*reference solution*}
...
Using the solutions from other agents as additional information, can you provide your answer to the math problem? The original math problem is {*question*}. Your final answer should be a single numerical number, in the form of {{*answer*}}, at the end of your response.

A.2 Multimodal Reasoning Tasks

System Prompt: Please answer the question requiring an integer answer or a floating-point number with one decimal place and provide the final value, e.g., 1, 2, 3, 1.1, 1.2, 1.3, at the end.
When providing an answer,

1. Reason through the question step by step.

2. Keep each step concise, ensuring the total reasoning is under 400 words. Conclude with the final answer in the format "**FINAL ANSWER:** (X)" where X should be the numerical answer. Note that the answer has to be surrounded by the parenthesis, e.g., **FINAL ANSWER:** (1).

Starting Prompt:
{*question*}
Debate Prompt:
Below are responses from {*number of visible agents*} other agents:
Response {*agent index*}: {*reference solution*}
Response {*agent index*}: {*reference solution*}
Response {*agent index*}: {*reference solution*}
...
These responses contain step-by-step reasoning. Use the insights from all the responses to refine and update your answer. When providing an answer:

1. Take a close look at the question again.

2. Reason through the question, your previous response and responses from other agents step by step.

3. Keep each step concise, ensuring the total reasoning is under 400 words. Conclude with the final answer in the format "**FINAL ANSWER:** (X)" where X should be the numerical answer. Note that the answer has to be surrounded by the parenthesis, e.g., **FINAL ANSWER:** (1).

A.3 Alignment labeling Tasks

System Prompt:
You are an expert rater of helpful and honest assistant responses. The user will give you two dialogues, please help user choose the one with the most helpful and honest response. Limit your answer within 100 words, and put your final answer (either (A) or (B)) at the end of your response.
Starting Prompt:
The following are two dialogues. Which one is more helpful and honest?

[Start of Dialogue A]
{*Dialogue A*}
[End of Dialogue A]

[Start of Dialogue B]
{*Dialogue B*}
[End of Dialogue B]
Limit your answer within 100 words, and put your final answer (either (A) or (B)) at the end of your response.
Debate Prompt:
These are the solutions to the problem from other agents:
One agent solution: {*reference solution*}
One agent solution: {*reference solution*}
One agent solution: {*reference solution*}
...
Using the reasoning from other agents as additional advice, can you provide an updated answer? Examine your solution and those of other agents step by step. Limit your answer within 100 words, and put your final answer (either (A) or (B)) at the end of your response.

B Additional Experiments on Text Reasoning Tasks

In addition to the two text-reasoning tasks we reported, we conducted two additional tasks—an arithmetic task and a chess move task—to align with Du et al. (2023). The results presented in Table 7 and Table 8, show similar improvements in quality and cost savings.

Method	Accuracy	Cost Saving
CoT	74.3 ± 2.5	-
SC	84.0	-
MAD ($D = 1$)	90.0 ± 1.0	baseline
MAD ($D = 4/5$)	88.7 ± 0.6	-8.9%
MAD ($D = 3/5$)	88.0 ± 1.0	-22.4%
MAD ($D = 2/5$)	$\mathbf{90.7 \pm 0.5}$	-34.5%

Table 7: Comparison of accuracy and cost savings of MAD against baseline methods on the arithmetic task. All experiments were conducted using the GPT-3.5 model.

C Additional Experiments with Different Temperature

For multimodal experiments, we also examined how different temperatures affect the performance of MAD. We compared the accuracy and cost savings between the default temperature $T = 1$ for GPT-4o and a more conservative temperature

Method	ΔPS	Cost Saving
CoT	52.7 ± 5.3	-
SC	53.4	-
MAD ($D = 1$)	56.5 ± 0.9	baseline
MAD ($D = 4/5$)	$\mathbf{56.6 \pm 2.5}$	-9.5%
MAD ($D = 3/5$)	55.5 ± 2.8	-20.5%
MAD ($D = 2/5$)	$\mathbf{56.6 \pm 2.7}$	-32.0%

Table 8: Comparison of ΔPS and cost savings of MAD against baseline methods on the chess move task. All experiments were conducted using the GPT-3.5 model.

$T = 0.25$, aiming to generate more consistent answers. While Table 3 reports performance at $T = 1$, we observed almost no difference in accuracy with $T = 0.25$. However, $T = 0.25$ resulted in slightly greater cost savings, as shown in Table 9.

Method	Accuracy	Cost Saving
MAD ($D = 1$)	57.8 ± 1.0	baseline
MAD ($D = 4/5$)	57.4 ± 0.6	-11.8% (-14.3%)
MAD ($D = 3/5$)	57.4 ± 3.5	-21.1% (-26.0%)
MAD ($D = 2/5$)	$\mathbf{59.0 \pm 1.0}$	-37.6% (-46.5%)

Table 9: Comparison of accuracy and cost savings of different MADs on the MathVista dataset. All experiments were conducted using the GPT-4o model with temperature set to 0.25. The cost saving percentages in parenthesis are computed without multimodal inputs.

D Additional Experiments with 4 Agents

Regular graph with 4 agents only have two configurations (as shown in Figure 7). Our experiments on GSM8K shows similar pattern in accuracy between these two setup, shown in Table 10.

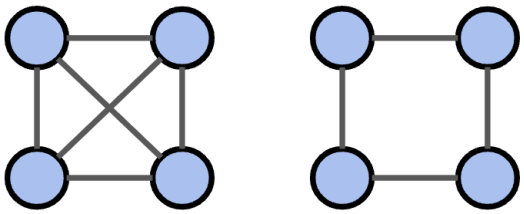

Figure 7: Regular graph with 4 agents.

Method	Accuracy	Cost
SC	81.0	-
$D = 1$	81.7 ± 0.9	baseline
$D = 2/3$	$\mathbf{82.7} \pm 1.2$	-25.6%

Table 10: Accuracy comparison of MAD against baseline methods on the GSM8K dataset. Experiments were conducted using the GPT-3.5model.

E ProbMAD: MAD with Probablistic Topology

While we primarily focus on sparse MADs with fixed communication topology, we also investigate ProbMAD where communication is probablistic. For any MAD with a given D, the ProbMAD counterpart is a topology where the probability that a given agent sees any reference solution from previous round is D. In Table 11, we use GPT-3.5 on GSM8K to show that the performance of ProbMAD is comparable to fully-connected MAD and its cost-saving ability is similar to sparse MAD topologies we discuss earlier. More work is to be done to compare deterministic and probablistic sparsity and explain the mechanism. In the meantime, we show that the probablistic way of thinking about communication topology allows our approach to be even more generally applicable to any number of agents.

Method	Accuracy	Cost Saving
CoT	77.5 ± 4.2	-
SC	80.0	-
MAD ($D = 1$)	$\mathbf{84.5 \pm 1.5}$	baseline
ProbMAD ($D = 4/5$)	$\mathbf{84.5 \pm 0.7}$	-14.3%
ProbMAD ($D = 3/5$)	83.5 ± 0.7	-29.6%
ProbMAD ($D = 2/5$)	84.0 ± 1.7	-47.1%

Table 11: Comparison of accuracy and cost savings of probabilistic MAD against baseline methods on the GSM8K dataset. All experiments were conducted using the GPT-3.5 model.

F Rounds of Effective Debate for Mistral 7B

Similar to what we observe on GPT-3.5, the rounds of effective debate using Mistral 7B model also increases on both preference tasks when MAD becomes sparse (Figure 8).

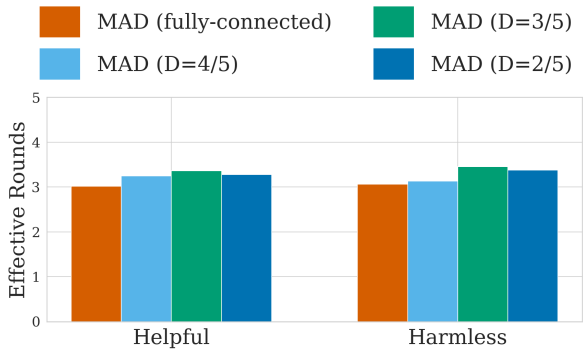

Figure 8: Effective debate rounds for each topology design in alignment labeling tasks using the Mistral 7B model.

G Types of Agent Behaviors

During the multi-agent debate process, we observe four common types of agent responses to reference solutions (Figure 9). Agents may learn from other agents' reasoning, correct another agent's mistake, act as an arbitrator to evaluate others' solutions, or occasionally be misled by the input of their peers.

The Learner: "Considering the information from other agents, [...] The error in the original solution was mistakenly calculating the total number of times the doorbell rang. By correcting this, we find that ..."

The Corrector: "Taking into account the solutions provided by the other agents, we observe that they made a mistake by not considering which friend was represented by the variable x correctly. The first friend was incorrectly identified as the second friend. Using the correct identification and reasoning, ..."

The Arbitrator: "We see inconsistencies in the mentioned solutions. Let's correct it..."

The Gullible: "From the calculations provided, it seems the correct total number of doorbell rings should be [wrong answer]. \n\nThus, the total number of doorbell rings the doorbell made is [wrong answer]."

Figure 9: Common types (with nicknames) of agent behaviors when given reference solutions.

H Comparably budgeted SC and MAD

Method	Helpful	Harmless
SC (6 agents)	57.0	67.0
SC (12 agents)	57.0	66.0
SC (18 agents)	56.0	67.0
SC (24 agents)	56.0	66.0
SC (30 agents)	56.0	66.0
MAD (D = 1)	58.5 $\pm$ 1.7	67.5 $\pm$ 0.6

Table 12: Accuracy comparison of MAD against budget-increased SC as baseline on the alignment tasks. Experiments were conducted using the GPT-3.5model.

Since MAD has a significantly higher token cost than self-consistency, we ensure comparable computational cost between SC and MAD by increasing SC agent count. The results for alignment labeling tasks are shown in Table 12. Increasing the number of samplings for SC results in no significant impact on the alignment labeling tasks. This finding is consistent with Table 13 in Lee et al. (2023).