GroupDebate: Enhancing the Efficiency of Multi-Agent Debate Using Group Discussion

Tongxuan Liu
University of Science and Technology of China
Hefei, CHINA
& JD.com
Beijing, CHINA
tongxuan.ltx@mail.ustc.edu.cn

Xingyu Wang
Institute of Automation, Chinese Academy of Sciences
Beijing, CHINA
wangxingyu2024@ia.ac.cn

Weizhe Huang
JD.com
Beijing, CHINA
huangweizhe1@jd.com

Wenjiang Xu
Institute of Automation, Chinese Academy of Sciences
Beijing, CHINA
xuwenjiang2024@ia.ac.cn

Yuting Zeng
University of Science and Technology of China
Hefei, CHINA
yuting_zeng@mail.ustc.edu.cn

Lei Jiang
University of Science and Technology of China
Hefei, CHINA
jianglei0510@mail.ustc.edu.cn

Hailong Yang
Beihang University
Beijing, CHINA
hailong.yang@buaa.edu.cn

Jing Li
University of Science and Technology of China
Hefei, CHINA
lj@ustc.edu.cn

ABSTRACT

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse NLP tasks, including complex logical reasoning, mathematical problem-solving, and multi-step decision-making. Extensive research has explored how to enhance the logical reasoning abilities such as Chain-of-Thought, Chain-of-Thought with Self-Consistency, Tree-Of-Thoughts, and multi-agent debates. In the context of multi-agent debates, significant performance improvements can be achieved with an increasing number of agents and debate rounds. However, the escalation in the number of agents and debate rounds can drastically raise the tokens cost of debates, thereby limiting the scalability of the multi-agent debate technique. To better harness the advantages of multi-agent debates in logical reasoning tasks, this paper proposes a method to significantly reduce token cost in multi-agent debates. This approach involves dividing all agents into multiple debate groups, with agents engaging in debates within their respective groups and sharing interim debate results between groups. Comparative experiments across multiple datasets have demonstrated that this method can reduce the total tokens by up to 46.9% during debates and while potentially enhancing accuracy by as much as 21.9%. Our method significantly enhances the performance and efficiency of interactions in the multi-agent debate.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/DSDX8860

KEYWORDS

Large Language Models; Multi-Agent Debate; Collaborative Reasoning

ACM Reference Format:
Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2026. GroupDebate: Enhancing the Efficiency of Multi-Agent Debate Using Group Discussion. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/DSDX8860

1 INTRODUCTION

Large language Models (LLMs) such as GPT [1, 5, 6, 27, 28], LLaMa [33, 34], and PaLM [2, 8] have revolutionized the field of natural language processing (NLP) by demonstrating remarkable capabilities across a wide range of tasks. These models can reach or even exceed human performance in a range of NLP tasks but their performance is still limited in complex mathematical and logical reasoning tasks [24]. To address these limitations, researchers have proposed developed various techniques to enhance the reasoning abilities of LLMs. Chain-of-Thought [18, 26, 37] encourages models to generates the reasoning process step by step. Subsequent research has introduced such as Tree-of-Thoughts [40] and Verification [22] to enhance their ability to perform complex multi-step reasoning. Unfortunately, these single-agent methods are more likely to fall into random fabrication of facts or the generation of delusions, thus leading to erroneous outcomes [6, 15, 16]. The multi-agent debate methods mitigate these issues by allowing different agents to express their arguments to each other and these approaches have demonstrated considerable potential and effectiveness across various types of tasks and datasets [7, 10, 21, 31, 35, 38, 39].

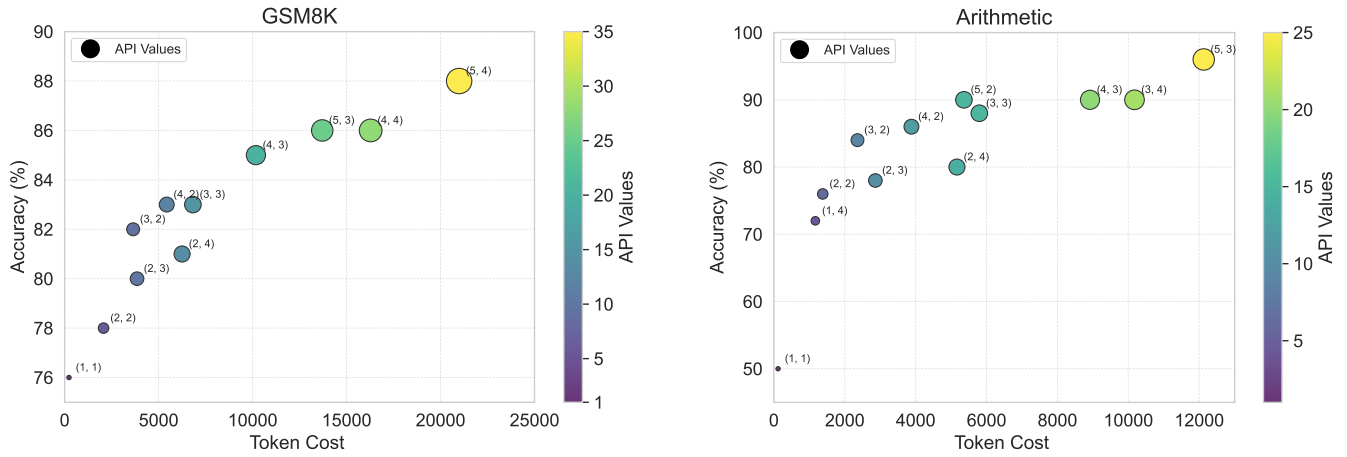

Figure 1: Comparison of Token Cost and Accuracy Under Different Combinations of Agents and Rounds. The numbers in parentheses corresponding to each circle represent the pair of agent number and round number. The size/color of the circle represents the number of API calls, indicating that the larger the circle, the more times the OpenAI API is called.

However, as the number of agents and rounds increases, the token cost in multi-agent debate can escalate significantly. This issue results in monetary expenditure on tokens through LLM-based API or substantial computational overhead and power consumption, thereby severely hindering the scalability and broader application of multi-agent debate, especially in scenarios with limited computational resources [12]. As illustrated in the Figure 1, compared with a single LLM-based agent, employing a multi-agent debate with three agents in five rounds can potentially raise the accuracy from the initial 50% to 98%, but introduces 101× token cost in the Arithmetic [5] task. Similarly, in the GSM8K [9] task, five rounds of multi-agent debate involving four agents can raise the accuracy from 76% to 88%, but it results in 90× token cost. To address the issue of the rapidly increasing number of tokens in multi-agent debates, researchers have proposed various improved techniques. For instance, the multi-agent debate in [10] summarizes the output of other agents to serve as the input for the next round. [31] proposes a "forgetfulness" mode that only the output from the previous round is stored as input for the next round. However, only employing a "forgetfulness" mode or summary mechanism to reduce token cost is still limited due to their theoretical complexity and the issue of exacerbated token growth. Moreover, owing to their simplistic debating modes, they struggle to fully exploit the collaborative capabilities of multi-agent debates.

In human societies, when multiple individuals engage in a debate, they usually conduct group discussion to enhance the efficiency of interaction while also preserving the diversity of viewpoints [19]. Inspired by this, in this paper, we propose a novel method GroupDebate (GD), which leverages group discussion to further reduce token cost in multi-agent debates. Specifically, Our method divides all participating agents into several debate groups, with each group conducting internal debates. Following the debates, the results are summarized and placed into a shared pool. After that, each group of agents retrieves the debate summaries of all groups from the pool, which serve as the input for the agents in the next round. Upon the conclusion of the debate, all agents reach a consensus or the final outcome is determined by majority vote. Furthermore, we conduct a theoretical analysis of the total token cost of the GroupDebate, thereby affirming the effectiveness of the method. In our experiments, we evaluate the effectiveness of GroupDebate in comparison to existing multi-agent debate methods and observe up to 34.8%/45.2%/46.9%/39.3%/30.6% reduction in token cost in the Arithmetic/GSM8K/MMLU/MATH/GPQA dataset, as well as up to 12%/21.9% improvement in accuracy in the MMLU/GPQA dataset.

Our main contributions are as follows:

1. We propose an innovative multi-agent debate strategy GroupDebate based on group discussion which can improve the efficiency and performance of multi-agent debates.
2. We conduct a theoretical analysis of token cost based on our method, demonstrating its efficiency and effectiveness.
3. Extensive experiments across five logical reasoning and mathematical datasets show that our method can not only significantly reduce token cost but also potentially enhance accuracy, validating the efficiency and superiority of our method.

2 PRELIMINARIES

2.1 Multi-agent Debate

In the context of multi-agent debates (MAD), by integrating multiple LLMs (each treated as an individual agent) and using various collaboration strategies, agents can propose viewpoints, review, and respond to the results of other agents in multiple rounds of debates [7, 31, 32]. The process of MAD can be summarized as follows:

(i) At the beginning, each agent is provided with a question and generates an individual response;

(ii) These responses then form the new input context for each agent, and the agents generate new responses;

(iii) This debate procedure is repeated over multiple rounds and the final answer is obtained through majority voting.

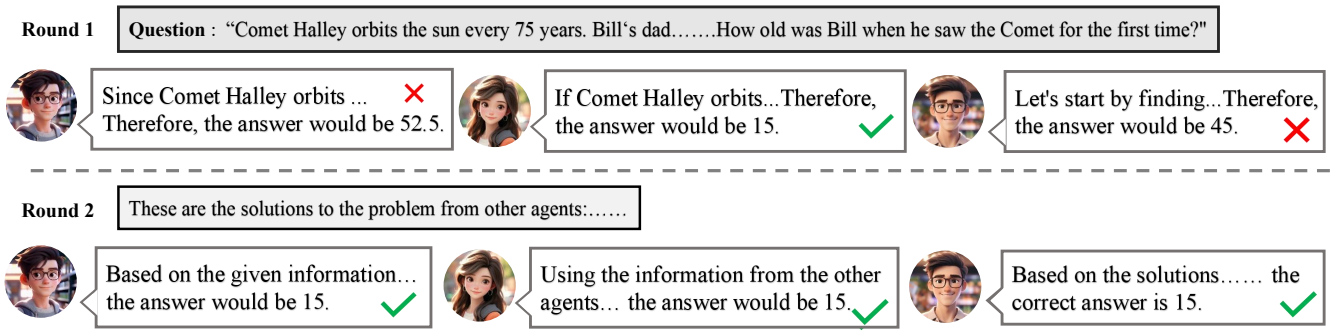

Figure 2: An Example of Multi-agent Debate Among Three Agents with Two Rounds.

Throughout multi-agent debate procedure, all agents can consistently improve their own responses based on the responses of other agents. In order to reduce input context length, [10] proposes that after collecting the responses from other agents, the responses should first be summarized and then used as the new input context for each agent. Figure 2 shows an example of two-round debates among three agents. In the first round, each agent independently responds to the input and their outputs are collected and summarized. In the second round, each agent's input includes summaries from the previous round, which are combined with a prompt to guide the output. Ultimately, all agents reach a consensus conclusion.

2.2 Token Cost in Multi-agent Debate

In the Figure 1, we can observe that although an increase in the number of agents and rounds can significantly enhance accuracy, the sharply increasing token cost is still a serious challenge in multi-agent debate. We further analyze this token cost issue based on the Simultaneous-Talk interaction strategy [7], where each agent synchronizes their results with other agents in each round of the debate. From Figure 3, it can be observed that under 4 rounds, as the number of agents increases from 1 to 8, the token cost in GSM8K/Arithmetic/MMLU has respectively grown by 36×/44×/49×. Similarly, under 4 agents, as the number of rounds increases from 1 to 4, the token cost in GSM8K/Arithmetic//MMLU has respectively increased by 17×/29×/19×.

3 METHODOLOGY

In this section, we first introduce the overall framework of our GroupDebate. Subsequently, we provide mathematical analysis of the token cost for both MAD and our GroupDebate. Formally, assume there are M LLM-based agents, denoted as $A = \{A_i | i = 1, 2, \cdots, M\}$, participating in a multi-round debate, with the total number of debate rounds denoted as T. In each round t ($t = 1, 2, \ldots, T$), the output of each agent A_i is represented as $Output_i^t$. These outputs are dynamically refined through structured inter-agent interactions, where agents critique, verify, and build upon reasoning from others. The tokens of the initial question prompt are denoted as Q, serves as the foundational query propagated through the debate process. These notations will be used throughout this paper.

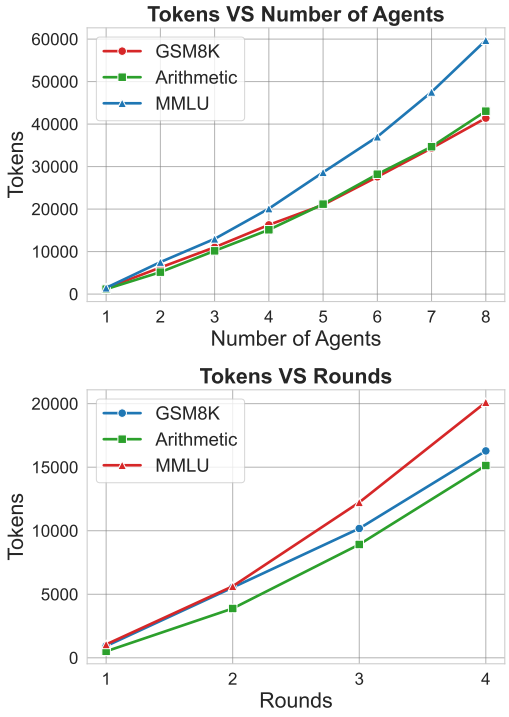

Figure 3: Token Cost Under Different Numbers of Agents and Rounds. The upper figure illustrates the token cost with variations in agents under the premise of 4 rounds. The lower figure illustrates the token cost with changes in rounds under the condition of 4 agents.

3.1 GroupDebate

The GroupDebate framework orchestrates collaborative reasoning among M LLM-based agents, denoted as $A = \{A_i | i = 1, 2, \cdots, M\}$, which can be randomly divided into N groups $G = \{G_j | j = 1, 2, \cdots, N\}$, with average K agents in each group. The GroupDebate splits the total debate rounds into S stages, with each stage encompassing R rounds. Thus, the total number of rounds T can be calculated as $T = S \times R$. This hierarchical design enables efficient exploration of solution spaces through alternating phases of localized refinement (intra-group) and global synthesis (inter-group). For the

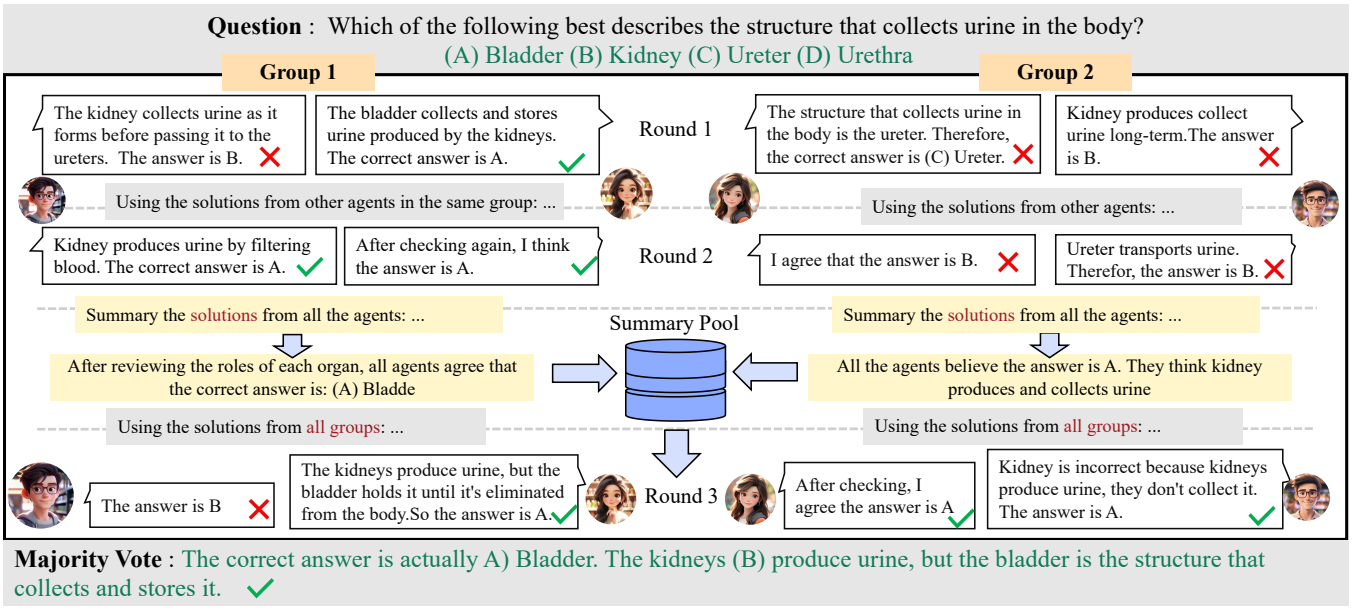

Figure 4: An Example of GroupDebate. 4 agents are divided into 2 groups and the GroupDebate process comprises two stages, with each stage involving two rounds of intra-group debate.

s-th stage's r-th round, GroupDebate selects one of the following processes:

(1) **Initial Thinking.** If $s = 1$ and $r = 1$ (i.e., the first stage's first round), we input the initial question prompt Q to each agent.
(2) **Intra-group Debate.** If $r > 1$, we utilize the outputs from other agents within the same group as the input for each agent.
(3) **Inter-group Debate.** If $s > 1$ and $r = 1$, we merge the outputs from the last round of each group into a summary and input the summaries from other groups to each agent.

Meanwhile, inspired by [31], we summarize the responses from other groups and restrict each agent to receive the latest summary from the previous stage in the inter-group debate. After the S-th stage's R-th round, all agents vote, and the ultimate result is determined by the majority selection. The Figure 4 illustrates an example of GroupDebate consisting of two stages and two groups. In the first stage, two agents in each group receive the initial question and exchange ideas within the group. In the second stage, agents share the summaries of their respective groups between groups and then discuss within their own groups again.

3.2 Token Cost Analysis

Token Cost in Multi-agent Debate. We implement the summary mechanism in MAD following [10], where the output of other agents is summarized and used as input for each agent in the next round. The summary for agent A_i in round t is denoted as $Summary_i^t$. Then token cost $Token^t$ in each round t can be computed as follows:

$$\begin{cases} \sum_{i=1}^{M}(Q + Output_i^t), & t = 1 \\ Token^{t-1} + \sum_{i=1}^{M}(S_i^{t-1} + Output_i^t), & t > 1 \end{cases} \tag{1}$$

Finally, the total token cost in MAD is

$$Token^{GD} = O\left(MTQ + (M^2T + MT^2)C\right) \tag{2}$$

where C represents the upper bound on the token number for each agent's response and the generated summary.

Token Cost in GroupDebate. In GroupDebate, we summarize the outputs from other groups at the end of each stage. Here, we define the summary of group G_j at the end of stage s as $Summary_j^s$.

Then token cost $Token_s^t$ in round t at stage s is:

Then the token cost in the intra-group debate is:

$$\sum_{j=1}^{N}\sum_{i\in G_j}(Q + Output_i^t + \sum_{i'\in G_j} Output_{i'}^{t-1}), \tag{3}$$

where $(s-1)R + 1 < t <= \min(sR, T)$. The token cost in the inter-group debate is:

$$\sum_{i=1}^{M}(Q + Output_i^{t-1} + \sum_{j=1}^{N} Summary_j^{s-1} + Output_i^t), \tag{4}$$

where $t = (s-1)R+1$. Finally, the total token cost of GroupDebate is $Token^{GD} = O\left(MTQ + (\frac{M^2T}{N} + MSN)C\right)$, where C represents the upper bound on the token number for each agent's response and the generated summary.

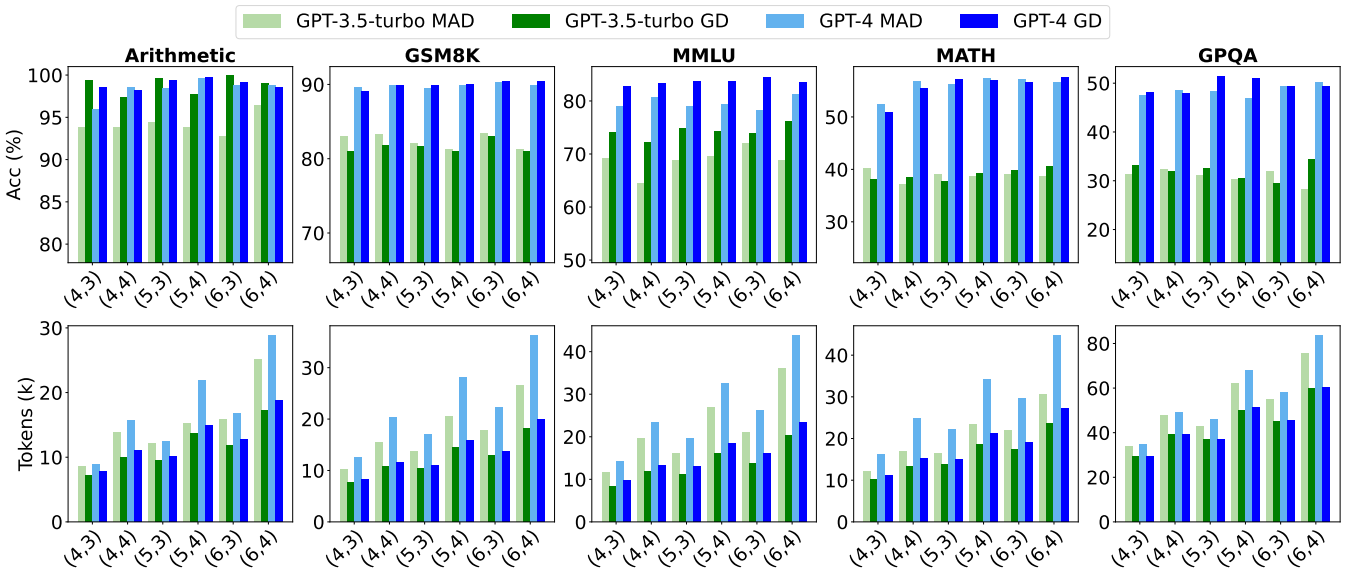

Figure 5: Comparison of Token Cost and Accuracy Between GD and MAD under Different Agents and Rounds. The notation (5,4) signifies 5 agents with 4 rounds. The results are average.

Discussion. From the perspective of overall token cost complexity, GD and MAD exhibit the same level of complexity regarding the input token cost of the question prompt Q, indicating that the question prompt has an equal impact on both methods. In our GroupDebate, given fixed values for T and M, the number of groups N and the total number of stages S can be dynamically adjusted. When we set $N \to O\left(\sqrt{\frac{MT}{S}}\right)$, theoretically, we can obtain $Token^{GD} \to O\left(MTQ + \sqrt{M^3TSC}\right)$. This complexity is significantly lower than that of MAD. If we consider setting S to a small positive integer, treating it as a constant, then $Token^{GD}$ can further approach $O\left(MTQ + \sqrt{M^3TC}\right)$. Moreover, N and S also influence the diversity in multi-agent debate, affecting the accuracy of the debate results, which will be further studied in Section 4.3.

4 EXPERIMENTS

4.1 Experimental Setup

Tasks and Metrics. To demonstrate the accuracy and effectiveness of different methods, we adopt total token cost and accuracy (ACC) as evaluation metrics. Additionally, we select five representative tasks related to logical reasoning and mathematical tasks to evaluate our methods, namely Arithmetic [5], GSM8K [9], MMLU [13], MATH [14] and GPQA [29].

Baselines. To evaluate the performance of GroupDebate (GD), We conduct a comparison of the efficiency and accuracy between GD and the following methods: (1) Chain-of-Thought (CoT) [37], which employs a sequential reasoning process to generate solutions; (2) Self-Consistency with Chain-of-Thought (CoT-SC) [36], an enhanced version of CoT that aggregates multiple reasoning paths to improve robustness, where CoT-SC(40) specifically denotes the use of 40 reasoning paths; (3) multi-agent debate (MAD) [21], a collaborative approach where multiple agents engage in iterative discussions to refine their solutions. For MAD, we explore various configurations by varying the number of agents and debate rounds. For example, both GD(5,3) and MAD(5,3) indicate configurations utilizing 5 agents and 3 rounds, allowing for a direct comparison of their performance under identical experimental conditions.

Implementation Details. We set the number of rounds of intra-group debate to 2 in GD. Additionally, we only retain output from the last round or summary generated from the last stage. Our experiments are conducted using the following models: `GPT-3.5-turbo-0301`, `GPT-4-0613`, `DeepSeek-R1-Distill-Qwen-32B`, `DeepSeek-R1` and `DeepSeek-V3`. In order to prevent the input prompt token exceeding the context limit, the MAD defaults to using the summary [10]. For all baselines and GD, we conduct five independent experiments separately and calculate the average. We evaluate these methods in a zero-shot setting.

4.2 Main Results

In this section, we present a detailed comparison of GroupDebate (GD) with multi-agent debate (MAD) and other single-agent methods, including Chain-of-Thought (CoT) and Self-Consistency with Chain-of-Thought (CoT-SC(40)). Notably, in the MATH dataset, MAD fails to produce results in both the (6,3) and (6,4) configurations due to the input prompt tokens exceeding the context limit of GPT-3.5. The key observations from our experiments are as follows:

Comparison Between GD and MAD.. First, as illustrated in Figure 5, GD consistently reduces token cost across different models under different agent and round settings, especially achieving up to 34.8%/45.2%/46.9%/39.3%/30.6% reduction in token cost in the Arithmetic/GSM8K/MMLU/MATH/GPQA datasets. This demonstrates

Table 1: Comparison of Token Cost and Accuracy Between GD and Other Methods. The results of highest accuracy are bold and the results of both highest accuracy and lowest token cost except from CoT are underlined

Methods	GPQA		GSM8k		Arithmetic		MMLU		Math	
	ACC(%)↑	Tokens(k)↓	ACC(%)↑	Tokens(k)↓	ACC(%)↑	Tokens(k)↓	ACC(%)↑	Tokens(k)↓	ACC(%)↑	Tokens(k)↓
`GPT-3.5-turbo-0125`										
CoT	31.2±0.03	2.0±0.02	76.8±0.02	0.25±0.00	82.2±0.04	0.16±0.02	70.2±0.02	0.24±0.00	35.2±0.02	0.37±0.01
CoT-SC(40)	31.2±0.02	80.5±0.26	**83.6**±0.01	10.0±0.02	95.0±0.01	6.3±0.11	**75.1**±0.02	**9.6**±0.03	**48.2**±0.01	14.6±0.15
MAD(5,3)	31.2±0.05	42.7±0.79	82.0±0.01	13.7±0.15	94.4±0.01	12.1±0.27	68.8±0.01	16.0±0.10	39.0±0.02	16.5±0.23
GD(5,3)	**32.6**±0.05	36.9±0.46	81.6±0.01	10.3±0.05	**99.6**±0.00	9.6±0.06	74.9±0.03	11.1±0.07	37.8±0.00	13.7±0.10
`GPT-4-0613`										
CoT	50.2±0.02	2.0±0.01	88.8±0.01	0.25±0.00	94.4±0.02	0.20±0.00	75.1±0.02	0.28±0.00	46.4±0.05	0.39±0.01
CoT-SC(40)	47.6±0.02	80.8±0.18	**91.0**±0.00	10.0±0.01	**100.0**±0.00	8.3±0.03	78.8±0.02	5.7±0.08	**62.8**±0.00	15.3±0.12
MAD(5,3)	48.4±0.02	46.1±2.07	89.4±0.00	17.1±0.06	98.4±0.01	12.5±0.01	79.0±0.02	20.0±0.11	56.2±0.02	22.3±0.24
GD(5,3)	**51.4**±0.03	36.9±1.11	89.8±0.02	11.0±0.01	99.4±0.01	10.1±0.33	**83.7**±0.01	13.0±0.12	57.2±0.03	15.1±0.19
`DeepSeek-R1-Distill-Qwen-32B`										
CoT	30.3±0.02	3.27±0.57	90.7±0.02	0.58±0.00	99.0±0.01	0.55±0.00	79.9±0.01	0.90±0.02	56.3±0.03	1.62±0.02
CoT-SC(40)	57.7 ±0.02	143.1 ±1.51	92.7±0.00	23.6±0.12	**100**±0.00	22.4±0.04	83.3±0.00	36.5±0.01	71.7±0.03	64.5±0.09
MAD(5,3)	**60.7**±0.02	125.5±2.12	92.7±0.01	57.9±0.28	100±0.00	65.8±0.50	84.0±0.02	66.5±0.52	87.0±0.02	104.8±0.24
GD(5,3)	59.0±0.01	78.4±3.23	**93.3**±0.01	31.9±0.48	100±0.00	33.5±0.33	**85.4**±0.01	31.9±0.21	**89.0**±0.01	58.1±0.58
`DeepSeek-R1`										
CoT	39.3±0.00	3.80±0.00	93.2±0.01	0.94±0.00	95.2±0.00	0.73±0.01	84.0±0.01	1.04±0.01	69.2±0.01	1.64±0.02
CoT-SC(40)	61.7 ±0.02	140.2 ±1.21	94.2±0.00	37.6±0.21	**100**±0.00	22.4±0.04	88.2±0.01	42.6±0.01	78.2±0.01	65.4±0.04
MAD(5,3)	62.0±0.01	148.2±1.13	**95.2**±0.02	61.7±0.23	99.0±0.01	55.3±0.30	86.7±0.02	81.4±0.12	90.2±0.02	84.5±0.12
GD(5,3)	**63.0**±0.01	72.9±0.02	94.4±0.01	26.6±0.28	99.4±0.00	12.9±0.13	**90.2**±0.02	34.4±0.01	**90.2**±0.01	47.9±0.08
`DeepSeek-V3`										
CoT	50.7±0.00	2.15±1.3	95.2±0.02	0.36±0.01	100±0.01	0.25±0.00	86.3±0.01	0.41±0.02	76.4±0.03	0.84±0.02
CoT-SC(40)	50.0 ±0.04	87.1 ±6.4	94.2±0.02	37.6±0.42	**100**±0.00	9.88±0.04	**88.4**±0.01	16.4±0.12	84.4±0.03	33.2±0.39
MAD(5,3)	54.0±0.02	74.8±2.37	**95.6** ±0.01	33.0±0.26	100±0.00	29.1±0.20	86.7±0.01	42.0±0.22	85.2±0.12	33.0±0.14
GD(5,3)	**54.0**±0.02	45.6±1.23	94.4±0.01	13.8±0.31	100±0.00	10.4±0.13	87.3±0.01	17.3±0.04	**86.0**±0.01	24.4±0.12

that our method can effectively reduce token cost in multi-agent debate while being theoretically grounded. Second, GD also improves accuracy in most settings, achieving up to 4.6%/4.9%/1.2% improvement in accuracy in certain settings of the Arithmetic/MMLU/GPQA dataset, which suggests GD can potentially enhance accuracy in multi-agent debate while reducing much token cost.

Comparison Between GD and Other Methods. As shown in Table 1, GD(5,3) and MAD(5,3) can significantly outperforms the standard single-agent method CoT, demonstrating the superiority of multi-agent debate in terms of accuracy. Moreover, multi-agent methods generally incur higher token cost compared to single-agent methods, indicating a significant challenge in reducing token cost among them. Our GD method can achieve the lowest token cost while ensuring high accuracy among all methods except CoT. Besides, we observe that CoT-SC(40) can achieve comparable performance of both accuracy and token cost with GD in average.

Table 2: Comparison of different group strategies within GD(6,4) on the MMLU dataset. The notation 3+3 denotes two groups, each consisting of three agents. The best results are bold.

Strategy	ACC(%)↑	Tokens(k)↓
MAD(6,4)	68.8±0.02	36.0±0.23
3 + 3	**76.1**±0.03	20.3±0.17
2 + 2 + 2	74.7±0.02	**18.6**±0.09
4 + 2	75.3±0.01	21.3±0.05

4.3 In-Depth Analysis

Group Strategy. In order to investigate the impact of different group strategy on accuracy and token cost, we conduct comparative experiments using GD(6,4) and `GPT-3.5-turbo-0125` on the

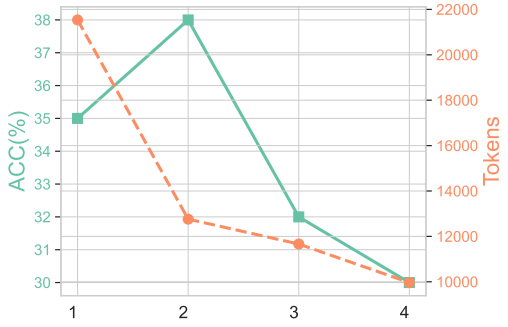

Figure 6: Different Intra-group Debate Rounds. The variations in accuracy are brought about by different intra-group rounds R.

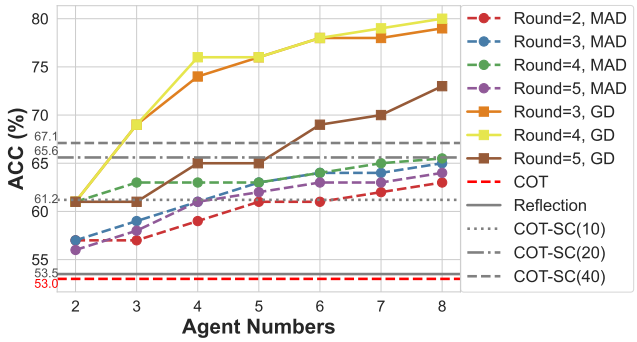

Figure 7: Scaling Study of Agents and Rounds.

MMLU dataset. As illustrated in Table 2, different group strategies can consistently enhance accuracy performance and reduce token cost compared to not grouping, demonstrating the effectiveness of group discussion. Moreover, 3+3 achieves the best accuracy in our experiment, indicating that more groups do not always mean the better accuracy performance. We leave the exploration of optimal group strategy parameters for achieving the best accuracy to future work.

Intra-group Debate Rounds. To explore the impact of the number of intra-group debate rounds, we conduct analysis under the condition of 4 agents and 4 rounds with varying numbers of intra-group debate rounds. As shown in Figure 6, best accuracy can be achieved when the number of intra-group debate rounds R is 2. This suggests that brief intra-group discussion can achieve better accuracy. Moreover, as R increases, the number of stages S decreases, resulting in lower token cost, which aligns with our derived complexity formula.

4.4 Scaling Study

Agent and Round Scaling. In order to explore the influence of rounds and agents on accuracy under MAD and GD, we evaluate the changing trends of accuracy for MAD and GD under various rounds and agents. As shown in Figure 7, with the increase in rounds, there is a significant growth in accuracy, but when rounds exceeds 4, a decrease in accuracy is observed across different numbers of agents. This reflects the phenomenon that limited increase in rounds can enhance accuracy, but excessive debate rounds can lead to accuracy degradation. As the number of agents increases, there is a significant growth in accuracy, indicating that an increase in agents can notably enhance the accuracy for both MAD and GD. Concurrently, it should be noted that the rate of improvement in accuracy tends to gradually decelerate as the number of agents continues to rise. The experimental results indicate the importance of controlling the appropriate number of agents and rounds.

Token Scaling. We assess the scaling trends of token cost and accuracy under both MAD and GD through by increasing the number of rounds or agents. First, as illustrated in Figure 8, with the increase in token cost, both MAD and GD exhibit an overall upward trend in accuracy. And initially the accuracy increases rapidly, but as the token cost becomes very large, the rate of accuracy growth slows down. Moreover, GD consistently outperforms MAD with scaling of tokens across all four datasets. While MAD's accuracy tends to converge as the token cost becomes exceedingly large, GD still potentially exhibits a growing trend. And we notice that GD has more sharply increasing points, which may be indicative of emergent intelligence in the token scaling of GD. It's an intriguing research point to explore scaling laws about accuracy and efficiency within multi-agent debate.

4.5 Ablation Study

In order to further investigate the impact of certain components in GD, we conduct a comparative analysis of MAD, MAD+Forget (MAD with only preserving summaries from the previous round), MAD+Group (MAD with group discussion) and GD. First, as illustrated in the Figure 9, GD outperforms all MAD and its variants in token cost and accuracy, which shows the effectiveness of involving both forget mechanism and group discussion in our method. Second, through comparing MAD+Forget with MAD and GD with MAD+Group, the forget mechanism can effectively reduce token cost while maintaining accuracy almost unchanged, which suggests that there is no need for agents to remember all summary results. Third, MAD+Group, compared to MAD+Forget, reduces a substantial number of tokens and significantly improves accuracy. This further highlights the effectiveness of our proposed group discussion method. Based on the grouping strategy analyzed previously, we hypothesize that the primary reason for the enhancement in accuracy is due to the diversity preserved among the groups.

5 RELATED WORK

5.1 LLM Reasoning

Numerous research have explored to enhance the logical reasoning capabilities of LLMs. Chain-of-Thought [37] is a pioneering work that mirrors human thought processes in a step-by-step way when tackling complex problems. Self-Consistency with CoT [36] samples multiple reasoning path and selects the most consistent answer. Tree-of-Thoughts [40] allows LLMs to determine their next course of action by considering various reasoning paths and self-evaluation choices. Graph-of-Thoughts [4] further represents the nonlinear task resolution process of LLMs as an arbitrary graph and

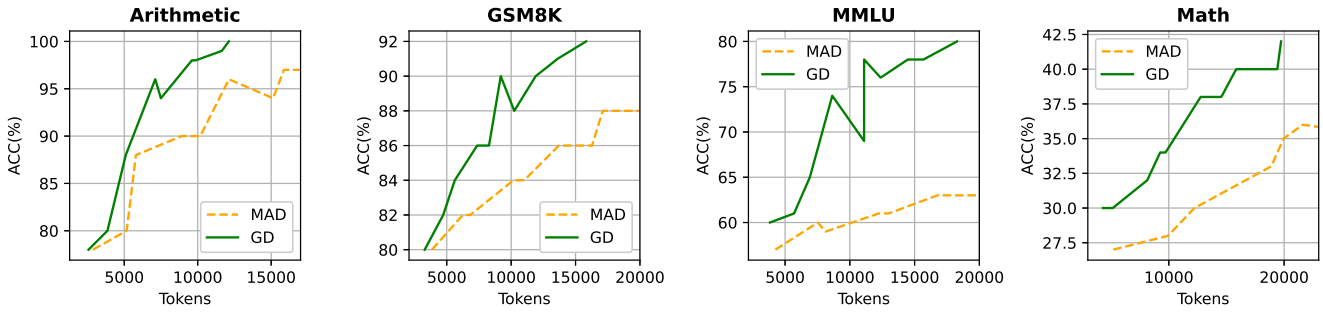

Figure 8: Scaling Study of Token Cost.

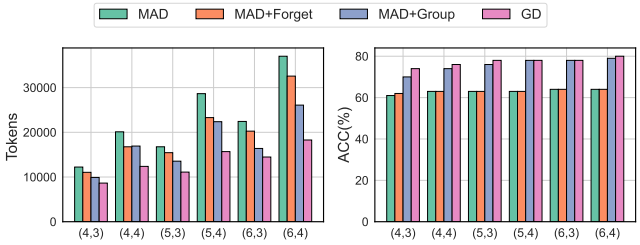

Figure 9: Ablation Study.

reasoning on the graph. Additionally, [30] involves creating a pool of CoT candidates and selecting the optimal candidate based on certain conditions. However, these methods either use only a single agent or lack communication between agents, which makes them prone to hallucinations or self-perception errors. Verification [22] and feedback recording are used to enhancement reasoning capabilities. STaR [41] generates multiple chains of thought, from which effective ones are selected. [42] proposes a method for selecting the optimal prompt from the candidate set. Skeleton-of-Thought [25] firstly generates skeleton of answer, followed by the parallel complete of content for each point in the skeleton, thus accelerating answer generation. Table-of-Thoughts [17] enhances the accuracy of reasoning through the structured modeling of the reasoning process.

5.2 Multi-agent Debate

In multi-agent collaboration, the MAD approach has been demonstrated as an effective orthogonal enhancement in logical reasoning. [21] proposes a MAD framework that encourages divergent thinking in LLMs, where a judge manages the debate and obtain a final solution. [10] further investigates the impact of the number of agents and rounds of debate on accuracy. [39] proposes a multi-agent collaboration strategy that simulates the academic peer review process. [35] integrates a prior knowledge retrieval into the debate process, thereby enhancing reasoning capabilities. [11] employs autonomous enhancement of negotiation strategies using a multi-round negotiation game exploration model with two agents. [7] presents various communication strategies and evaluates the effects of these differing approaches. Corex [31] employs collaborative methods such as debate, review, and retrieve among multiple agents. [20] proposes that the utilization of a sparse communication topology in MAD to enhance performance and mitigate computational costs. Recent studies [3] have identified the critical issue of problem drift – a gradual deviation from the original task objective. ConfMAD [23] integrates confidence expression throughout the debate process to improve effectiveness andperformance.

6 CONCLUSION

In this work, we address the critical challenge of token efficiency in multi-agent debate systems, which has emerged as a key bottleneck in scaling collaborative reasoning frameworks. We propose a novel GroupDebate method, which leverages the group discussion to mitigate this issue while fostering a diverse range of viewpoints. Specifically, we divide all participating agents into several debate groups, where each agent can engage in both intra-group debates and inter-group exchanges of ideas. Experimental results across four logical reasoning datasets demonstrate GroupDebate can significantly reduce token cost as well as enhance accuracy in multi-agent debates. In the future, we will further explore the theorem of how group discussion can improve accuracy and theoretically determine the optimal settings in GroupDebate.

7 LIMITATIONS

While GroupDebate demonstrates significant advancements in token efficiency and reasoning accuracy, several limitations and avenues for future research remain: The first key limitation is that we only theoretically analyze the constraints of N and S required to achieve optimal token cost complexity but we have not delved into the optimal settings of N and S and the underlying reasons why GroupDebate can potentially improve the accuracy of MAD. However, determining the optimal values of N and S also requires considering accuracy to maximize it under the same token cost, which is very complex and necessitates more evaluations and experiments to deduce the theoretical basis for the enhancement of accuracy and optimal settings in GroupDebate. Furthermore, although GroupDebate can significantly reduce token cost in muti-agent debates, its token cost is still higher than single-agent methods like CoT. It is necessary to explore more ways to further reduce token cost while ensuring high accuracy.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (62322201).

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[3] Jonas Becker, Lars Benedikt Kaesberg, Andreas Stephan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Stay Focused: Problem Drift in Multi-Agent Debate. arXiv:2502.19559 [cs.CL] https://arxiv.org/abs/2502.19559

[4] 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*, Vol. 38. 17682–17690.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[7] 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. *arXiv preprint arXiv:2308.07201* (2023).

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[9] 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. *arXiv preprint arXiv:2110.14168* (2021).

[10] 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* (2023).

[11] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142* (2023).

[12] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. arXiv:2402.01680 [cs.CL]

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[14] 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. *arXiv preprint arXiv:2103.03874* (2021).

[15] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).

[16] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[17] Ziqi Jin and Wei Lu. 2023. Tab-cot: Zero-shot tabular chain of thought. *arXiv preprint arXiv:2305.17812* (2023).

[18] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[19] Richard A Krueger. 2014. *Focus groups: A practical guide for applied research*. Sage publications.

[20] Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving Multi-Agent Debate with Sparse Communication Topology. *arXiv preprint arXiv:2406.11776* (2024).

[21] 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* (2023).

[22] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050* (2023).

[23] Zijie Lin and Bryan Hooi. 2025. Enhancing Multi-Agent Debate System Performance via Confidence Expression. arXiv:2509.14034 [cs.CL] https://arxiv.org/abs/2509.14034

[24] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439* (2023).

[25] Xuefei Ning, Zinan Lin, Zixuan Zhou, Huazhong Yang, and Yu Wang. 2023. Skeleton-of-thought: Large language models can do parallel decoding. *arXiv preprint arXiv:2307.15337* (2023).

[26] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114* (2021).

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[29] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022 [cs.AI] https://arxiv.org/abs/2311.12022

[30] KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822* (2023).

[31] Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2023. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *arXiv preprint arXiv:2310.00280* (2023).

[32] Mikhail Terekhov, Romain Graux, Eduardo Neville, Denis Rosset, and Gabin Kolly. 2023. Second-order Jailbreaks: Generative Agents Successfully Manipulate Through an Intermediary. In *Multi-Agent Security Workshop@ NeurIPS'23*.

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[35] Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2023. Apollo's Oracle: Retrieval-Augmented Reasoning in Multi-Agent Debates. *arXiv preprint arXiv:2312.04854* (2023).

[36] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[37] 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 (2022), 24824–24837.

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

[39] Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023. Towards reasoning in large language models via multi-agent peer review collaboration. *arXiv preprint arXiv:2311.08152* (2023).

[40] 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 (2024).

[41] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems* 35 (2022), 15476–15488.

[42] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022).