Debate Only When Necessary: Adaptive Multiagent Collaboration for Efficient LLM Reasoning

Anonymous ACL submission

Abstract

Multiagent collaboration has emerged as a promising framework for enhancing the reasoning capabilities of large language models (LLMs). Despite improvements in reasoning, the approach introduces substantial computational overhead resulting from iterative agent interactions. Furthermore, engaging in unnecessary debates increases the risk of generating erroneous responses. To address these challenges, we propose Debate Only When Necessary (DOWN), an adaptive multiagent collaboration framework that integrates a deterministic gating module conditioned on the initial response. Debate is activated exclusively for queries that necessitate further deliberation, wherein agents refine their outputs by leveraging peer responses and their associated confidence scores. Evaluations on benchmarks show that DOWN improves efficiency by up to *six times* while preserving or even outperforming the performance of existing methods. Further analysis indicates that DOWN effectively mitigates the risk of error propagation stemming from the redundant debate process. These findings demonstrate the effectiveness of our approach in delivering high-performance LLM solutions at a lower computational cost.

1 Introduction

Building on the remarkable advancements in large language models (LLMs), recent research has increasingly focused on extending their capabilities to address complex real-world problems (Yao et al., 2023; Fan et al., 2024; Chen et al., 2024b). Among various research directions, multiagent collaboration has emerged as a promising approach, inspired by human decision-making processes in complex problem-solving (Minsky, 1988; Li et al., 2023; Chen et al., 2024a; Wang et al., 2025; Wu et al., 2023; Du et al., 2024). By engaging in structured debate, LLM agents systematically exchange perspectives and iteratively cross-examine each other's reasoning to refine their responses. This collaborative process facilitates divergent thinking and enhances the reasoning capabilities of LLMs (Liang et al., 2024; Chen et al., 2024b; Chan et al., 2024).

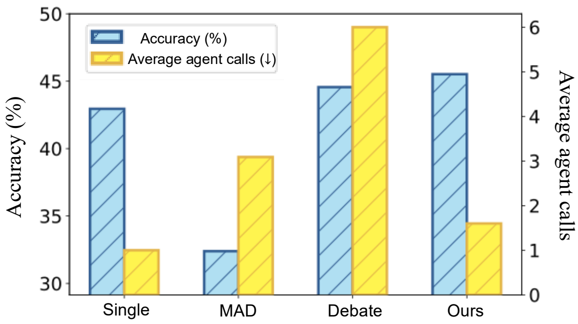


Figure 1: Comparison of accuracy and average agent calls across various multiagent debate methods

Despite these advantages, multiagent collaboration systems exhibit several key limitations. From an efficiency perspective, iterative interactions among agents inherently require multiple agent calls, leading to increased latency and higher inference costs (Snell et al., 2024; Kapoor et al., 2024). Meanwhile, when agents engage in redundant debate, there is a higher likelihood of generating errors that may subsequently be propagated by other agents in the system (Wang et al., 2024). Figure 1 illustrates both of these issues by plotting the accuracy and average agent calls of different multiagent debate methods. In the Debate system, additional debate rounds improve performance while incurring a sixfold increase in computational overhead. In the MAD framework, despite using more agent calls than the single agent baseline, its accuracy paradoxically declines. Regarding the practical application of multiagent collaboration systems, these challenges emphasize the need for an optimized collaboration approach (Kapoor et al., 2024; Tran et al., 2025).

To address these limitations, we propose Debate

Only When Necessary (DOWN), an adaptive multiagent collaboration framework designed to activate debate selectively. At the core of DOWN is a deterministic gating module that operates solely on the initial agent response and determines, on the basis of its associated confidence score, whether further debate is necessary. The gating module is designed with the objective of minimizing computational cost while preserving performance and is optimized through a utility score that balances efficiency and predictive accuracy. This conditional activation allows the system to invoke debate only for queries that benefit from deeper collaborative reasoning, reducing computation without sacrificing quality. Upon debate activation, agents engage in a confidence-guided collaboration, drawing on their peers' contributions and integrating the most credible and compelling elements to produce a more accurate and coherent final outcome.

We employ models of varying scales, including approximately 8B and 70B parameter models, evaluated on the MUSR, StrategyQA, CommonsenseQA, and MMLU benchmarks. The findings demonstrate that adaptive debate engagement achieves up to a sixfold reduction in computational overhead while maintaining or even surpassing the performance of full-debate baselines. Notably, we reveal that this conditional debate serves as a safeguard against cascading errors, effectively enhancing the advantages of agent collaboration. This work offers the following principal contributions:

- We propose Debate Only When Necessary (DOWN), an adaptive multiagent framework that selectively initiates debate based on the initial output, achieving up to a sixfold cost savings on representative reasoning benchmarks while preserving competitive performance.

- Extensive analysis reveals the strong robustness and generalization capability of our approach, substantiating its broad applicability. Moreover, the adaptive gating mechanism is found to effectively mitigate error propagation.

- To the best of our knowledge, this is the first study to investigate the gating mechanism in multiagent systems grounded in the initial response, maximizing efficiency while preserving the benefits of multiagent collaboration.

2 Related Work

2.1 LLM-based Multiagent Collaboration

Drawing inspiration from human collaborative problem-solving behavior, multiagent collaboration systems leverage collective intelligence to improve decision-making. Studies have demonstrated that LLM-powered multiagent systems promote divergent thinking (Xiong et al., 2023; Liu et al., 2024; Liang et al., 2024) and improve reasoning capabilities (Li et al., 2023; Yin et al., 2023; Zhuge et al., 2023). With these advantages, multiagent collaboration is leveraged for diverse NLP applications: mitigating hallucinations (Fang et al., 2025), aggregating knowledge across multiple specialized LLMs (Wang et al., 2025), generating novel scientific ideas and insights (Su et al., 2024), evaluating LLM-generated responses (Chan et al., 2024), and refining datasets for instruction fine-tuning (Li et al., 2024a). These advancements highlight the growing impact of multiagent collaboration.

2.2 Debate Structures in Multiagent Collaboration Systems

Recent studies have developed debate structures to optimize the benefits of multiagent collaboration. For instance, Du et al. (2024) introduces a framework in which agents iteratively refine their responses based on peer-generated outputs. Liang et al. (2024) propose a structured debate format that assigns distinct roles to encourage divergent thinking. Wang et al. (2024) develop a discussion system in which agents are organized into multiple groups to engage in discussions.

However, iterative debate frameworks face a fundamental trade-off between efficiency and accuracy (Kapoor et al., 2024; Snell et al., 2024). To this end, Li et al. (2024b) restrict discussions to local neighbors, while Zhou et al. (2025) and Liu et al. (2024) introduce a shortcut mechanism that shares similarities with our approach. However, both methods still incur non-trivial agent calls. The former necessitates non-trivial interactions among neighboring agents, and the latter relies on a self-consistency criterion that requires at least five agent calls per sample. Additionally, iterative debates may propagate errors or introduce a trade-off between diversity and quality (Wang et al., 2024; Kapoor et al., 2024; Li et al., 2025). Our approach focuses on these limitations by proposing a resource-efficient and performance-effective debate framework.

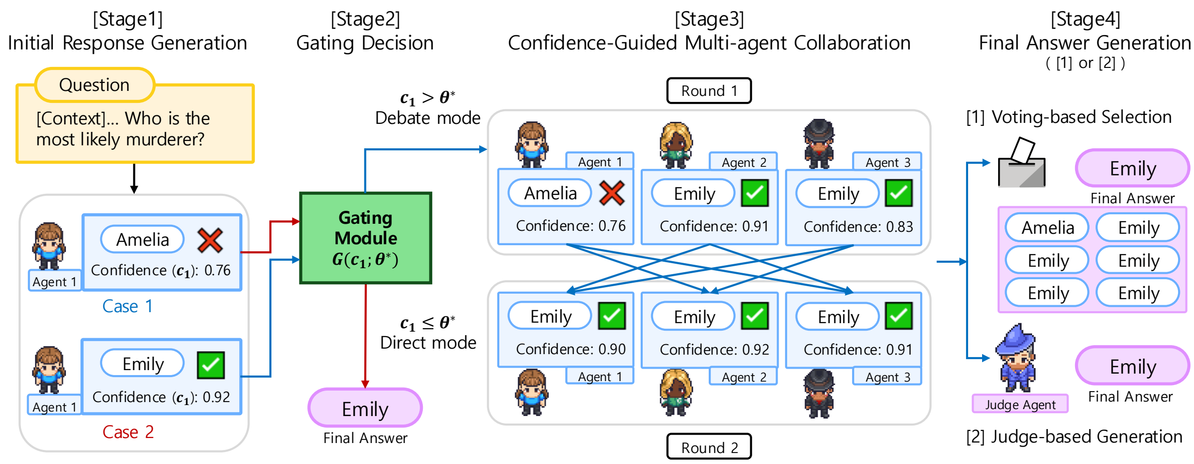


Figure 2: The DOWN framework begins with an initial agent generating a response, after which the gating module determines whether further collaboration is required based on the confidence score of the initial output c_1. Queries deemed acceptable ($G(c_1; \theta^*) = 0$) proceed via the direct mode without invoking additional agents, whereas those considered uncertain ($G(c_1; \theta^*) = 1$) trigger a collaborative process in which agents iteratively refine their answers by referencing the outputs and confidence scores of their peers. The final answer is subsequently determined either through majority voting or by a designated judge agent.

3 Debate Only When Necessary

We introduce Debate Only When Necessary (DOWN), an adaptive multiagent collaboration framework that engages in debate based on the gating module. This section provides a detailed description of the debate procedure and the optimization of the gating module. Figure 2 illustrates the overall architecture.

3.1 System Pipeline

Stage 1: Initial Response Generation Given an input query q, an initial agent $\mathcal{A}_1$ from a pool of agents $\{\mathcal{A}\}_{i=1}^{M}$ generates an initial response r_1, accompanied by a reason supporting its answer. We explore two configurations to ensure broad applicability: (1) A homogeneous-agent configuration where all collaborating agents use the same model (e.g., Llama) and (2) a heterogeneous-agent configuration where the initial agent is randomly selected for each query (e.g., Llama, Qwen, GPT-4o-mini). Along with the response, a confidence score is extracted to quantify the certainty of the prediction, a commonly used measure in prior work (Razghandi et al., 2025; Taubenfeld et al., 2025; Chen et al., 2024a). We first obtain the token logits $L(t_i)$, where $t_i \in r_1$, from the hidden representation of the generated response. These logits are then passed through a softmax function to yield the probability distribution $P(t_i)$ over tokens. To obtain a robust estimate of confidence, we apply normalization by averaging the token probabilities over the generated response. The confidence score c_1, is defined as:

$$c_1 = \frac{1}{|r_1|}\sum_{i=1}^{|r_1|} P(t_i), P(t_i) = \frac{\exp(L(t_i))}{\sum_j \exp(L(t_j))}. \tag{1}$$

For models without access to internal logits, we employ verbalized confidence as a proxy.

Stage 2: Gating Decision The gating module determines whether to engage in the collaborative debate guided by the initial response's confidence score c_1, which is formalized as:

$$G(c_1; \theta) = \begin{cases} 0, & \text{if} \quad c_1 > \theta \quad \text{(Direct mode)}, \\ 1, & \text{if} \quad c_1 \le \theta \quad \text{(Debate mode)}, \end{cases} \tag{2}$$

where θ is the gating parameter, and its optimization procedure is described in the subsequent subsection. When the gating function assigns a query to $G(c_1; \theta) = 0$, the system directly accepts the initial response as the final output without invoking further agents, reducing computational overhead. Conversely, when $G(c_1; \theta) = 1$, the query is routed to the debate process, enabling agents to collaboratively refine and enhance the response.

Stage 3: Confidence-Guided Multiagent Collaboration The collaboration process involves response generation and refinement. In our experiments, we set up an environment with two rounds, each consisting of three agents. To conduct multiagent collaboration, in round 1, we obtain responses

r_2 and r_3 from additional agents $\mathcal{A}_2$ and $\mathcal{A}_3$, along with their respective confidence scores c_2 and c_3. These confidence scores are explicitly concatenated with each response to convey the agent's certainty level in subsequent rounds.

After all agents have completed their first-round responses, the collaboration proceeds to the next round. Each agent $\mathcal{A}_i$ gains access to the responses produced by the other agents while excluding its own, denoted as $\mathcal{R}_{-i} = \{r_j | j \neq i\}$. Leveraging this additional context $\mathcal{R}_{-i}$, each agent refines its reasoning and generates an updated second-round response $\{r_i^{(2)}\}_{i=1}^3$, capitalizing on more confident and persuasive arguments.

Stage 4: Final Answer Generation The responses generated in Stage 3 serve as ingredients for deriving the final answer. We explore two distinct strategies for final answer output: voting-based selection and judge-based generation. The voting-based approach determines the final answer by selecting the most frequent response among all agent-generated outputs. We design this majority voting to enhance robustness by leveraging consensus across multiple perspectives, effectively mitigating individual agent biases. The judge-based approach introduces an additional judging agent, which generates the final response based on all agent outputs. We induce a judging mechanism to update the final response, prioritizing well-supported and coherent arguments.

3.2 Gating Module Optimization

The performance of DOWN hinges on properly designing the gating module. We optimize the gating parameter θ by optimizing a utility-driven objective that controls the trade-off between predictive fidelity and computational efficiency.

Balancing Quality and Efficiency Let $A_\theta = \frac{k}{N}$ and $S_\theta = \frac{m}{N}$ denote the accuracy and debate skip rate (i.e. efficiency) for a given threshold value θ in the discrete candidate set Θ. To prevent sacrificing quality for efficiency, we identify a high-performance region using the one-sided 95% Wilson lower bound of the maximum observed accuracy. Given the highest observed accuracy A_θ^{max} obtained from k^{max} correct predictions, the Wilson lower bound $\tilde{A}_\theta^{max}$ is computed as:

$$\tilde{A}_\theta^{\max} = \frac{2k^{\max} + z^2 - z\sqrt{z^2 + 4k^{\max}\left(1 - \frac{k^{\max}}{N}\right)}}{2(N + z^2)}, \tag{3}$$

where $z = \Phi^{-1}(0.95) \approx 1.645$. Threshold candidates whose accuracy lies within this region are considered statistically indistinguishable from the best-performing value, whereas those outside the region $A_\theta < \tilde{A}_\theta^{\max}$ are subject to soft penalization.

In parallel, the design prevents convergence to degenerate behaviors, such as engaging in debate either almost constantly or almost never. A penalty applies whenever the skip rate lies outside the interval $[s_{\min},\ 1-s_{\min}]$, where $s_{\min}$ denotes the proportion symmetrically excluded from the lower and upper ends of the unit interval $[0, 1]$. The total penalty term is defined as: $P_\theta = \max(0, \tilde{A}_\theta^{\max} - A_\theta) + \max(0, s_{\min} - S_\theta) + \max(0, S_\theta - (1 - s_{\min}))$.

Utility-based Optimization Min–max scaling is employed for normalization, balancing the contributions of the two components to the final score:

$$\hat{A}_\theta = \frac{A_\theta - A_\theta^{\min}}{A_\theta^{\max} - A_\theta^{\min} + \epsilon}, \hat{S}_\theta = \frac{S_\theta - S_\theta^{\min}}{S_\theta^{\max} - S_\theta^{\min} + \epsilon}, \tag{4}$$

where $\epsilon \ll 1$ ensures numerical stability. The gating parameter is then optimized by maximizing the utility score: $\theta^* = \arg\max_{\theta\in\Theta}(\hat{A}_\theta + \hat{S}_\theta - \lambda \cdot P_\theta)$, where λ controls the relative strength of the penalty.

At inference time, the optimized gate module $G(c_1; \theta^*)$ governs the processing path. Queries with acceptable outputs proceed through the direct mode, whereas uncertain cases are directed to the debate mode. Though optimized, the gating module further permits flexible adjustment to user preferences, with higher thresholds favoring accuracy and lower thresholds emphasizing efficiency.

4 Experiments

Experimental Setup The evaluation considers two agent configurations: homogeneous and heterogeneous. The homogeneous setting employs a single model across all agents, using Llama-3.1 8B, Ministral 8B, Qwen-2.5 72B, Llama-3.3 70B, and GPT-4o-mini. The heterogeneous-model configuration combines Qwen-2.5 72B, Llama-3.3 70B, and GPT-4o-mini, with the model selection order randomized for each query. Each debate round involves three agents, and the debate process consists of two rounds, where initial responses are generated in the first round and refined in the second. Our experiments are conducted on three benchmarks of MUSR (Sprague et al., 2024), StrategyQA (Geva et al., 2021), and CommonsenseQA (CSQA) (Talmor et al., 2019), with MMLU (Hendrycks et al.,

Benchmark	Method	Ministral 7B		Llama3.1-7B		GPT-4o-mini		Llama-3.3-70B		Qwen-2.5-70B		Average	
		Acc	AC (↓)	Acc	AC(↓)	Acc	AC(↓)	Acc	AC(↓)	Acc	AC(↓)	Acc	AC(↓)
MUSR	Single CoT	51.06	1.00	42.95	1.00	55.75	1.00	56.33	1.00	57.80	1.00	52.78	1.00
	Self-refine	36.90	9.00	39.46	9.00	54.29	9.00	53.67	9.00	58.47	9.00	48.56	9.00
	Self-Consistency	48.24	6.00	44.70	6.00	55.88	6.00	**58.18**	6.00	58.29	6.00	53.06	6.00
	MAD	28.67	3.01	32.39	3.09	43.23	3.02	51.22	3.00	49.13	3.04	40.93	3.03
	Debate	48.54	6.00	44.56	6.00	57.32	6.00	57.28	6.00	58.69	6.00	53.28	6.00
	DOWN (Vote)	**53.71**	**1.48**	45.51	**1.50**	57.09	**1.80**	57.80	**1.02**	59.39	**1.28**	54.70	**1.42**
	DOWN (Judge)	**53.71**	1.57	**45.52**	1.60	**57.35**	1.96	57.80	1.03	**59.52**	1.34	**54.78**	1.50
StrategyQA	Single CoT	67.69	1.00	70.74	1.00	78.17	1.00	80.35	1.00	78.60	1.00	75.11	1.00
	Self-refine	67.69	9.00	69.54	9.00	76.42	9.00	77.73	9.00	78.17	9.00	73.91	9.00
	Self-Consistency	68.12	6.00	68.56	6.00	79.48	6.00	80.79	6.00	77.29	6.00	74.85	6.00
	MAD	57.64	3.73	44.54	4.66	70.31	3.38	79.04	3.07	73.80	3.24	65.07	3.62
	Debate	**70.74**	6.00	70.08	6.00	79.04	6.00	80.35	6.00	**79.91**	6.00	76.02	6.00
	DOWN (Vote)	68.12	**3.16**	**71.18**	**2.53**	**80.79**	**1.92**	**82.53**	**1.07**	77.73	**2.64**	**76.07**	**2.26**
	DOWN (Judge)	68.12	3.59	69.87	2.83	79.91	2.10	**82.53**	1.08	77.73	2.97	75.63	2.51
CSQA	Single CoT	70.60	1.00	70.02	1.00	82.15	1.00	83.70	1.00	84.60	1.00	78.21	1.00
	Self-refine	70.84	9.00	70.02	9.00	82.72	9.00	83.78	9.00	82.39	9.00	77.95	9.00
	Self-Consistency	69.78	6.00	**73.55**	6.00	82.39	6.00	85.18	6.00	**85.42**	6.00	**79.26**	6.00
	MAD	42.42	4.30	19.66	4.68	78.54	**3.00**	78.30	**3.00**	74.20	3.00	58.62	3.60
	Debate	67.73	6.00	70.52	6.00	**83.05**	6.00	**85.42**	6.00	84.60	6.00	78.26	6.00
	DOWN (Vote)	**71.33**	**4.11**	70.76	**1.59**	82.96	3.70	84.68	3.80	84.93	**2.27**	78.93	**3.09**
	DOWN (Judge)	71.17	4.73	71.33	1.71	82.80	4.24	84.68	4.36	85.01	2.52	79.00	3.51

Table 1: Comparison of accuracy (Acc) and average agent calls (AC) across single model strategies, multiagent debate systems, and our proposed approach on three benchmarks

2021) employed for further analysis of generalization performance. Accuracy is used as an evaluation metric, while efficiency is measured in terms of the average number of agent calls (AC) and debate skip rate. Further details of the models, benchmarks, and prompts are provided in Appendix A and Appendix B.

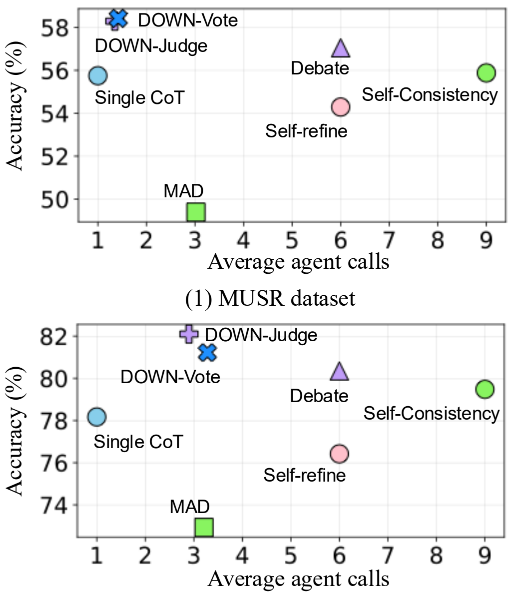
Baselines We conduct comparisons against multiple established reasoning systems, including a single Chain-of-Thought (CoT) (Kojima et al., 2022) baseline. (1) Self-Refine (Madaan et al., 2023): Each round consists of generating an initial response, receiving feedback, and producing a refined response. (2) Self-Consistency (Wang et al., 2023): Sampling responses multiple times and deriving the final answer by selecting the most consistent one through majority voting. (3) MAD (Liang et al., 2024): Conducting a debate between two agents with opposing perspectives while a moderator selects the most plausible solution or continues the debate if needed. (4) Debate (Du et al., 2024): Agents engage in a structured debate, iteratively updating their own responses by incorporating insights derived from other agents' responses in the previous round.

We additionally introduce baselines that incorporate various efficient collaboration strategies. For a fair comparison, each proposed efficiency mechanism is integrated into the standard Debate or MAD frameworks. The Early Stopping (Liu et al., 2024) terminates the debate when two out of three agent outputs are consistent in the initial round. Similarly, the Efficiency Optimization (Zhou et al., 2025) applies a consistency-based criterion that halts the debate when all of the outputs are consistent. Both methods utilize output consistency as the decision signal and are directly applied to the Debate method. In contrast, SparseMAD (Li et al., 2024b) improves efficiency by structuring debates through local neighbor interactions instead of fully connected discussions, which we incorporate into the MAD framework.

5 Results and Analysis

5.1 Main Results

Homogeneous-Model Configuration Table 1 presents the results on three benchmark datasets. Remarkably, our framework attains performance on par with or even outperforms single and multiagent baselines across the MUSR, StrategyQA, and CSQA benchmarks, while requiring only 1.42 to 3.09 AC on average, resulting in up to a sixfold improvement in efficiency. The framework demonstrates notable robustness across the Mistral, Llama, GPT, and Qwen families at parameter scales of 7B and 70B, attesting to the wide-reaching applicability of DOWN across diverse architectures. These findings highlight that conditional debate, rather



(2) StrategyQA dataset

Figure 3: Performance comparison under a heterogeneous-model configuration. For single model approaches, we report GPT-4o-mini results.

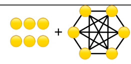
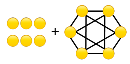
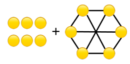
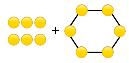
Method	Collaboration	Acc.	AC
Case1: Debate			
Early Stopping	[R1] [R2]	82.53	<u>5.04</u>
Efficiency Optimization	[R1] [R2]	<u>82.97</u>	10.00
DOWN	[R1] [R2]	**83.41**	**2.57**
Case2: MAD			
SparseMAD (D= 5/5)		79.04	21.00
SparseMAD (D= 4/5)		<u>79.91</u>	18.00
SparseMAD (D= 3/5)		**80.35**	15.00
SparseMAD (D= 2/5)		79.04	<u>12.00</u>
DOWN	[R1] [R2]	<u>79.91</u>	**3.01**

Table 2: Results comparing efficient debate strategies on the StrategyQA dataset using the Llama-70B model. Red and blue circles indicate dynamic debate control: red when debate is skipped because consistent responses reach the specified count, blue when skipped by the gating module. Yellow circles denote standard agent outputs. Each round is denoted by R, and the communication graph density by D.

than applying it uniformly or relying solely on a single agent response, offers a more principled balance between accuracy and efficiency. Interestingly, MAD exhibits lower average accuracy on all benchmarks. Consistent with findings from Wang et al. (2024), we attribute this to its inherent tendency toward contradictory reasoning. While constructive disagreement encourages divergent insights, it also intensifies erroneous reasoning, deteriorating the quality of final responses.

Heterogeneous-Model Configuration Under the heterogeneous-model configuration, external feedback originates from models trained on distinct datasets and scales, which facilitates the emergence of diverse and complementary reasoning behaviors. As shown in Figure 3, DOWN consistently outperforms strong baselines, achieving higher accuracy while demanding far fewer agent calls across both datasets. DOWN surpasses the Debate in performance, with a markedly reduced computational burden. These results demonstrate that DOWN remains effective beyond homogeneous setups, adapting seamlessly to heterogeneous-model collaboration. This highlights the robustness and scalability of our system in diverse reasoning environments.

Final Answer Generation Strategies Voting-based selection and judge-based evaluation are two strategies for consolidating debate responses into a final decision. Experimental results indicate that both approaches yield similar response patterns, with marginal differences depending on the specific setting. However, the judge-based method requires an additional agent call, making it slightly less efficient. Given this trade-off, the voting-based approach is preferable when prioritizing computational efficiency, as it achieves comparable accuracy with fewer computational resources.

5.2 Comparison to Efficient Debate Methods

Table 2 presents experimental results where existing efficiency strategies and our DOWN are integrated into the Debate (Case 1) and MAD (Case 2) frameworks. As a result, in Case 1, DOWN outperforms existing efficiency methods in both accuracy and efficiency. DOWN achieves the highest accuracy of 83.41 despite requiring only 2.57 agent calls. We attribute this to the fact that the Early Stopping and Efficiency Optimization require full agent response generation in the first round to evaluate consistency, limiting their efficiency gains. In Case 2, SparseMAD reduces communication density through neighbor-based interactions but continues to incur a substantial number of agent calls. DOWN, by contrast, attains comparable accu-

LLM Agent	Response Shift	MAD	Debate	Ours
GPT-4o-mini	✓→✗	70.59	50.00	**33.59**
	✗→✓	29.41	50.00	**66.41**
Llama-3.3 70B	✓→✗	48.91	60.09	**12.57**
	✗→✓	51.09	39.91	**87.43**
Qwen-2.5 72B	✓→✗	63.87	50.00	**39.91**
	✗→✓	36.13	50.00	**60.09**
Mix	✓→✗	70.85	**30.81**	47.35
	✗→✓	29.15	**69.19**	52.65

Table 3: Proportions of correct and incorrect response shifts induced by debate across multiagent collaboration systems. We denote a correct response as ✓and an incorrect response as ✗.

racy with only 3.01 agent calls, thereby matching the performance of the 4-out-of-5 (4/5) neighbor-connected configuration. These findings highlight the persistent reliance of existing efficient debate methods on a substantial number of agent calls. DOWN, by discerning the need for debate solely from the initial response, delivers markedly higher efficiency while sustaining robust performance.

5.3 Response Shifts in Accuracy

Table 3 reports the proportions of correct and incorrect changes in answers, computed over StrategyQA samples where the final answer differs from the initial prediction. To deepen our investigation, we set the threshold parameters to 0.9, increasing the number of cases where debate is triggered. Across all model configurations, DOWN consistently achieves a higher successful correction rate than baseline debate systems. Notably, for the Llama 70B model, 87.43% of changed answers reflect successful corrections of initially incorrect predictions. MAD shows a high rate of incorrect changes, erroneously revising correct initial answers into incorrect ones. The Debate baseline shows a similar rate of correct and incorrect changes, indicating a limited capacity to prioritize reliable peer inputs. The high success rate of corrections achieved by DOWN stems from two core design principles. It initiates debate when the gating module directs the query to debate mode, avoiding unnecessary revisions of correct answers. Once debate is engaged, it draws on peer responses with their associated confidence scores, enabling the model to integrate reliable information.

5.4 Ablation Study

To analyze the contribution of each component in the DOWN framework, we conduct an ablation study on the StrategyQA. The results are shown in Table 4. We begin by examining the role of confidence scores shared among agents during debate. Eliminating these signals results in a 3.06% point decline in accuracy, underscoring that confidence serves as a crucial cue for selectively incorporating peer responses throughout the debate process.

Method	Acc.	AC
DOWN (θ^*)	71.18	2.53
– Confidence-guided debate	68.12	2.46
– Gating module (single response only)	70.74	1.00
– Gating module (debate only)	70.74	6.00
+ Lowered gating params ($\theta^* - 0.1$)	69.87	1.20
+ Raised gating params ($\theta^* + 0.1$)	71.62	3.53

Table 4: Ablation study using the StrategyQA dataset

The subsequent ablation analysis scrutinizes the efficacy of the gating module. When the gating module is removed and the debate process is either entirely suppressed or universally enforced for all inputs, accuracy drops to 70.74 in both scenarios. This outcome highlights that the gating module mediates a principled balance between performance gains and computational cost by reducing the average number of agent calls. The findings further reveal that indiscriminate debate engagement is redundant, conferring no benefit over the single-agent baseline and indicating that existing multi-agent collaboration strategies remain suboptimal with respect to both accuracy and efficiency.

To further validate the efficacy of the gating parameter, we perturb θ^* by ± 0.1 and examine the resulting impact on performance. Deviations from the optimized value in either direction expose a trade-off between accuracy and the average number of agent calls. By contrast, the optimized threshold θ^* maintains accuracy within a statistically reliable range, as quantified by the Wilson lower bound, while markedly reducing computational cost.

5.5 Generalization Analysis

While DOWN has demonstrated strong effectiveness on reasoning tasks, we evaluate its generalization capability across diverse knowledge domains using the MMLU benchmark to substantiate its broader applicability. The performance of multiagent collaboration systems across six domains in MMLU is shown in Figure 5. Experimental results show that DOWN achieves performance comparable to that of the Debate method across tasks. Given that the Debate approach con-

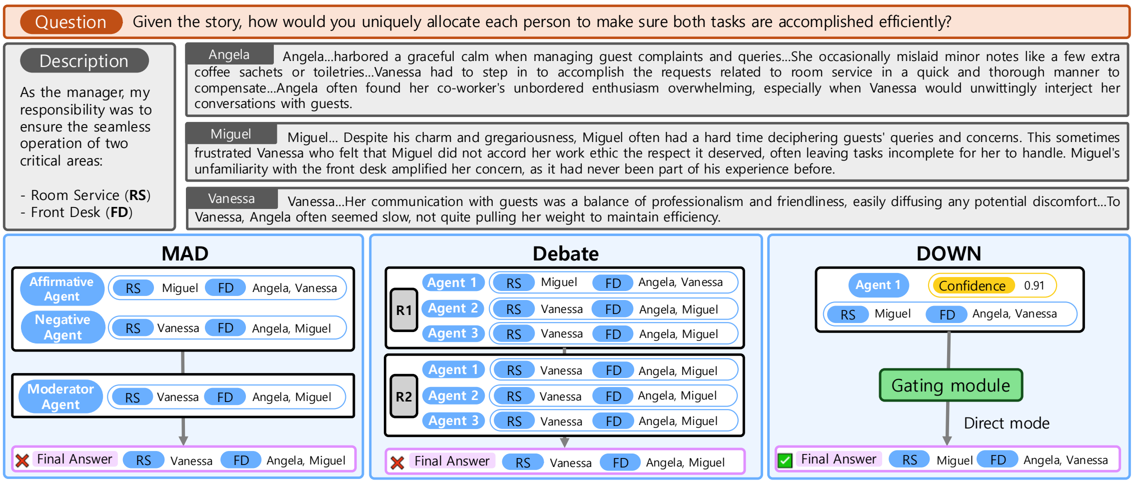


Figure 4: Qualitative comparison of multiagent collaboration methods evaluated on the MUSR dataset

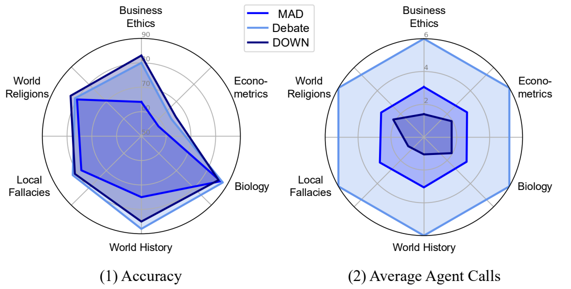


Figure 5: Accuracy and average agent calls (AC) of multiagent debate methods across six MMLU domains

sistently requires an average of six agent calls, the DOWN framework shows significantly higher efficiency. The results suggest that DOWN generalizes well, achieving robust performance on benchmarks evaluating both reasoning capabilities and factual knowledge across diverse domains.

From another perspective, *DOWN supports both dataset-specific construction and broadly applicable construction of the gating module, depending on the user preference*. As shown in Table 1 and Figure 5, gating modules produced under both settings consistently achieve strong and stable improvements in efficiency and accuracy. These results demonstrate that the gating module effectively accommodates a wide range of data conditions rather than overfitting to any particular dataset. This highlights DOWN's potential to serve as a scalable and effective alternative to full multiagent debate systems.

5.6 Qualitative Analysis

Figure 4 presents a qualitative comparison of responses generated by different debate systems on the MUSR dataset. The results show that existing debate mechanisms introduce unnecessary modifications. Although the initial responses produced by the Debate and MAD methods are correct, subsequent iterative revisions lead to an incorrect final prediction. This implies the potential risk of error propagation when redundant debate occurs. In contrast, with an initial confidence score of 0.91, DOWN skips the debate and directly adopts the initial response as the final answer. This suggests that the adaptive gating module under high-confidence conditions prevents unnecessary modifications while maintaining efficiency. To further substantiate these findings, we present additional qualitative analysis in the Appendix C.

6 Conclusion

This work proposed the DOWN framework to address the computational inefficiencies and error-propagation challenges inherent in multiagent collaboration. By conditionally activating debate based on the gating module, we significantly reduced computational overhead while maintaining or even improving performance. The results indicated that conditional debate improved efficiency and mitigated cascading errors, resulting in more stable and reliable reasoning dynamics. Furthermore, the confidence-guided multiagent debate amplified the influence of reliable responses on final decisions. Taken together, these findings position DOWN as a robust and principled optimization approach, delivering high-performance and computationally efficient solutions for multi-agent collaborative systems.

Limitations

While our proposed DOWN framework demonstrates strong efficiency and robustness, several limitations remain. While the confidence score widely serves as a fundamental indicator of a model's internal certainties (Razghandi et al., 2025; Taubenfeld et al., 2025; Chen et al., 2024a) and is adopted as the gating criterion in DOWN, it remains susceptible to over- or under-confidence. The use of model-specific optimized thresholds ensures consistent gating across models despite their varying degrees of over- or under-confidence, though further refinement through more sophisticated uncertainty estimation remains a promising direction for future work. This study primarily focuses on English debates, which may limit its applicability to multilingual multiagent collaboration settings. Extending our method to multilingual LLMs would provide deeper insights into its robustness across diverse linguistic contexts. Due to computational constraints, we employ the GPT-4o-mini instead of the GPT-4o model. Applying the GPT-4o model to our system could further deepen the understanding of our DOWN framework.

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024a. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024b. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–25.

Yi Fang, Moxin Li, Wenjie Wang, Lin Hui, and Fuli Feng. 2025. Counterfactual debating with preset stances for hallucination elimination of LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10554–10568, Abu Dhabi, UAE. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. Ai agents that matter. *arXiv preprint arXiv:2407.01502*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Renhao Li, Minghuan Tan, Derek F. Wong, and Min Yang. 2024a. CoEvol: Constructing better responses for instruction finetuning through multi-agent cooperation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4703–4721, Miami, Florida, USA. Association for Computational Linguistics.

Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. 2025. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024b. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024. A dynamic LLM-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Ali Razghandi, Seyed Mohammad Hadi Hosseini, and Mahdieh Soleymani Baghshah. 2025. Cer: Confidence enhanced reasoning in llms. *arXiv preprint arXiv:2502.14634*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *arXiv preprint arXiv:2410.09403*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590, Singapore. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153, Singapore. Association for Computational Linguistics.

Wei Zhou, Mohsen Mesgar, Annemarie Friedrich, and Heike Adel. 2025. Efficient multi-agent collaboration with tool use for online planning in complex table question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 945–968, Albuquerque, New Mexico. Association for Computational Linguistics.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

A Implementation Details

We utilize the pre-trained model parameters provided by HuggingFace, with the exact models listed as follows: Llama-3.1 8B (`meta-llama/Llama-3.1-8B-Instruct`), Mistral 8B (`mistralai/Mistral-8B-Instruct-2410`), Qwen-2.5 72B (`Qwen/Qwen2.5-72B-Instruct`), Llama-3.3 70B (`meta-llama/Llama-3.3-70B-Instruct`), and GPT-4o-mini (`gpt-4o-mini`). For the optimization of the gating module, we define the candidate set Θ to commence at 0.7 with an interval of 0.1, corresponding to a threshold that delivers a debate skip rate S_θ close to 1. The symmetric margin $s_{\min}$ and the penalty scaling factor λ are set to 0.1 and 15. For reproducibility and deterministic response generation, the temperature is set to 0.0, and the random seed is fixed at 777. The maximum generated sequence length is limited to 512 tokens. All experiments are conducted under consistent configurations across models using four A6000 GPUs with 48GB of memory.

For the evaluation, we utilize benchmarks designed to assess reasoning capabilities. MUSR (Sprague et al., 2024) evaluates multi-step soft reasoning over free-text narratives, offering a more complex yet realistic reasoning challenge compared to synthetic benchmarks. Assessments are performed on 756 samples drawn from the murder mysteries, object placements, and team allocation subsets. StrategyQA (Geva et al., 2021) requires implicit reasoning, where inference steps must be strategically derived rather than explicitly provided. Evaluations are conducted on the development set comprising 229 samples. CommonsenseQA (CSQA) (Talmor et al., 2019) is a multiple-choice dataset requiring commonsense reasoning beyond textual information. Evaluations are conducted on the development set of 1,221 samples. To further assess the generalization capability of the DOWN framework, we conduct additional experiments on the MMLU (Hendrycks et al., 2021) knowledge benchmark. We consider six domains of business ethics, econometrics, biology, world history, logical fallacies, and world religions that cover a broad spectrum of knowledge areas.

B Prompts leveraged in DOWN framework

The prompts employed in our experiments are presented in the following textboxes. These prompts are applied throughout the DOWN framework to ensure structured and consistent response generation.

Initial Answer Generation

[debate topic]
Please output your answer in json format, with the format as follows:
{"**base_reason**": "", "**base_answer**": ""}.
Please strictly output in JSON format, do not output irrelevant content.

Initial Answer Generation (w Confidence)

[debate topic]
Please output your answer in json format, with the format as follows:
{"**base_reason**": "", "**base_answer**": "", "**confidence_score**":range of 0-1}.
Please strictly output in JSON format, do not output irrelevant content.

Answer Update

These are the solutions to the problem from other agents:
One agent solution: **[agent1 response]**
One agent solution: **[agent2 response]**
Using the solutions from other agents as additional information, can you provide your answer to the problem?

Judge Prompt

Based on the following responses, generate an updated response that most accurately addresses the given query.
Query: **[debate topic]**
Responses: **[agent responses]**.
Summarize your reasons for selecting this side and provide the final answer.

C Qualitative Analysis of Debate Systems

Additional qualitative analysis for each multiagent collaboration method is presented in Table 5, Table 6, and Table 7.

Table 5 presents the responses for different debate methods on the StrategyQA dataset. While the initial response in the MAD system is correct, the system revises its answer after being influenced by the opposing stance of the negative debater. While this process encourages divergent thinking, it ultimately leads to incorrect conclusions due to error propagation. In contrast, the Debate methodology consistently generates correct responses across all cases. Although this approach required six agent calls, it enhances response reliability. The DOWN methodology, on the other hand, produces an initial response with a confidence score of 0.95, leading the system to skip the debate process. This result highlights the efficiency gains achieved through conditional debate.

Table 6 shows the results on the MUSR dataset. Experimental results indicate that both the MAD and Debate methodologies fail to conduct correct reasoning. The initial response in DOWN shows a confidence score of 0.89, while subsequent responses report 0.84 and 0.92, respectively. Notably, in the second round, the initial agent revised its response based on higher-confidence outputs from other agents, ultimately arriving at the correct answer. Model confidence scores also generally increase in the second round. This finding underscores the importance of confidence scores as a guiding metric, allowing the system to integrate reliable responses.

Table 7 follows a trend similar to Table 5. The DOWN system strategically skips the debate process, enhancing efficiency while maintaining answer accuracy.

Method	Agent	Category	Content
Query	Could a hundred thousand lolcats fit on a first generation iPhone?		
Answer	Yes		
MAD	Affirmative Debater	Reason	The storage capacity of a first generation iPhone is 4GB, 8GB, or 16GB, and the average size of an image file is approximately 100KB. ... However, ... it's theoretically possible for a hundred thousand lolcats to fit on a first generation iPhone with sufficient storage.
		Answer	Yes
	Negative Debater	Reason	No. The first generation iPhone had a maximum storage capacity of 16GB for the 16GB model, but also consider other factors such as the operating system, apps, and other data that would occupy some of that space. ... Considering these constraints, it's unlikely that a hundred thousand lolcats could fit on a first generation iPhone.
		Answer	No
	Moderator	Reason	While the affirmative side presents a theoretical calculation that a hundred thousand lolcats could fit within the storage capacity of a higher-end first generation iPhone, the negative side raises ... by the operating system, other data, and the file system. These factors significantly reduce the available space for storing images, making it unlikely that a hundred thousand lolcats could be accommodated.
		Answer	No
	Final Answer		No
Debate	Round 1 Agent 1	Reason	The first generation iPhone had a storage capacity of 4GB, 8GB, ... but exceeds the lower-end models. ... Assuming a very small size for each lolcat to fit on the screen, it's theoretically possible to display them, but it would be extremely impractical and unlikely.
		Answer	Yes
	Round 1 Agent 2	Reason	The first generation iPhone has a storage capacity of 4GB, 8GB, or 16GB, and assuming an average size of a lolcat image (around 100KB), we can estimate the maximum number of images that can fit. ... Still, theoretically, a hundred thousand lolcats could fit on a 16GB first-generation iPhone if we only consider storage space for images and nothing else.
		Answer	Yes
	Round 1 Agent 3	Reason	The first generation iPhone had a storage capacity of 4GB, 8GB, or 16GB, and assuming an average image size of 100KB, we can estimate the maximum number of images that can be stored. ... Nevertheless, from a purely storage-based perspective, it is theoretically possible for a hundred thousand lolcats to fit on a 16GB first generation iPhone.
		Answer	Yes

Method	Agent	Category	Content
	Round 2 Agent 1	Reason	Considering the storage capacity of the first generation iPhone, with models ranging from 4GB to 16GB, ... would require approximately 10GB of storage space. ... Given this interpretation and focusing solely on storage capacity, it is theoretically possible for a hundred thousand lolcats to fit on a 16GB first-generation iPhone, especially when considering only the storage aspect and disregarding the practicality of displaying them.
		Answer	Yes
	Round 2 Agent 2	Reason	Considering the storage capacity of the first generation iPhone, with models ranging from 4GB to 16GB, and estimating an average size of 100KB per lolcat image, a hundred thousand lolcats would require approximately 10GB of storage space. ... Therefore, focusing solely on storage, a hundred thousand lolcats could theoretically fit on a 16GB first-generation iPhone, albeit with the understanding that displaying them all at once would be highly impractical.
		Answer	Yes
	Round 2 Agent 3	Reason	Considering the storage capacity of the first generation iPhone, with models ranging from 4GB to 16GB, and assuming an average size of 100KB per lolcat image, a hundred thousand lolcats would require approximately 10GB of storage space. ... Taking into account the solutions from other agents, which also consider the theoretical possibility of storing a hundred thousand lolcats on a 16GB model, it appears that from a storage perspective, it is indeed possible.
		Answer	Yes
	Final Answer		Yes
DOWN	Round 1 Agent 1	Reason	The first generation iPhone had a storage capacity of 4GB, 8GB, or 16GB, and each lolcat image is approximately 100KB in size. A hundred thousand lolcats would require 10GB of storage space, which exceeds the capacity of the 4GB and 8GB models but fits within the 16GB model.
		Answer	Yes
		Confidence Score	0.95
	Final Answer		Yes

Table 5: Qualitative analysis of the StrategyQA dataset

<table>
<tr><th>Method</th><th>Agent</th><th>Category</th><th>Content</th></tr>
<tr><td>Query</td><td colspan="3">Given the story, how would you uniquely allocate each person to make sure both tasks are accomplished efficiently?

Choices: ‘Seeking Sponsors and Donations: Amelia, Organizing Event Details: George and Lily’, ‘Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily’, ‘Seeking Sponsors and Donations: Lily, Organizing Event Details: Amelia and George’

As the clock ticked closer to our most significant fundraising event, three key players - George, Amelia, and Lily - paced around the office, eager to know their assignments. As their manager, I held the responsibility of delegating two critical tasks among them: the challenging pursuit of Seeking Sponsors and Donations, and the intricate duty of Organizing Event Details.

Amelia, our vibrant member, often found herself with Lily, brainstorming creative ideas in the cafeteria or around the picnic table outside. ... She was infamous for her love for detailed schedules and checklists, which often found her volunteering to plan her family reunions that ran smoothly under her supervision. ...

However, the presence of George often marred her high spirits. George, indeed, was a hard one to deal with. Even though he had this endless passion ... , he often disagreed with Lily’s ideas during meetings, publicly criticized Amelia’s work enough to dampen her spirit, and worst of all, forgot crucial tasks that he was responsible for. ... But he often confessed feeling overwhelmed managing multiple tasks at once.

Then there was Lily, a networking maven. ... Yet, her true forte was her meticulousness that reflected in her attention to detail. ... While she tended to dismiss George’s inputs during meetings, she showed a great deal of respect to Amelia and considered her suggestions before making a decision.

Now, the ball was in my court. With the fundraising event nearing, it was my responsibility to ensure that these three unique individuals were correctly placed to utilize their strengths best. ... As the manager, I knew that this decision would shape the fate of our event and define our team’s dynamic in the foreseeable future.</td></tr>
<tr><td>Answer</td><td colspan="3">Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily</td></tr>
<tr><td rowspan="2">MAD</td><td rowspan="2">Affirmative Debater</td><td>Reason</td><td>Based on the characteristics and strengths of each individual, Amelia’s experience in attending charity galas and her ability to make connections, as well as her attention to detail, make her a strong candidate for Seeking Sponsors and Donations. ... George’s tendency to disagree with others and forget tasks suggests that he may not be the best fit for a role that requires independent work or a high level of organization, but his passion for fundraising can still be utilized in a team setting.</td></tr>
<tr><td>Answer</td><td>Seeking Sponsors and Donations: Amelia, Organizing Event Details: George and Lily</td></tr>
</table>

Method	Agent	Category	Content
	Negative Debater	Reason	I disagree with your answer. Lily's experience in sales and her ability to ask people for money, combined with her connections with wealthy friends and acquaintances, make her a strong candidate for Seeking Sponsors and Donations. ... It allows Lily to leverage her sales experience and connections to secure sponsors and donations, while Amelia and George work together to organize the event details, with Amelia's organizational skills helping to keep George on track.
		Answer	Seeking Sponsors and Donations: Lily, Organizing Event Details: Amelia and George
	Moderator	Reason	The negative side's allocation takes into account the strengths and weaknesses of each individual more effectively. ... George's passion for fundraising can still be utilized in a team setting, and his tendency to disagree with others and forget tasks is mitigated by working with Amelia and Lily.
		Answer	Seeking Sponsors and Donations: Lily, Organizing Event Details: Amelia and George
	Final Answer		Seeking Sponsors and Donations: Lily, Organizing Event Details: Amelia and George
Debate	Round 1 Agent 1	Reason	The task of Seeking Sponsors and Donations requires someone with strong connections and the ability to ask for money, which Lily possesses due to her past experience in sales and her network of wealthy friends and acquaintances. ... However, his experience in a marketing firm could still be valuable in a supporting role or in a different capacity, but for the two main tasks at hand, Amelia and Lily are the better fit.
		Answer	Seeking Sponsors and Donations: Lily, Organizing Event Details: Amelia and George
	Round 1 Agent 2	Reason	Based on the characteristics and interaction dynamics of the three individuals, Amelia's strengths in seeking sponsors and donations due to her connections and experience, as well as her attention to detail, make her a strong candidate for Seeking Sponsors and Donations. ... Instead, considering George's potential to disrupt the team's dynamic, it would be more efficient to have him focus on Seeking Sponsors and Donations, where his interactions are more limited, and utilize Amelia and Lily's collaborative and organized approach for Organizing Event Details.
		Answer	Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
	Round 1 Agent 3	Reason	To efficiently allocate tasks, we need to consider the strengths and weaknesses of each person. ... George, despite his flaws, has a passion for fundraising and experience in managing sponsorships, making him a better fit for Seeking Sponsors and Donations, where his skills can be utilized without disrupting the team's dynamics.
		Answer	Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily

Method	Agent	Category	Content
	Round 2 Agent 1	Reason	Considering the characteristics and interaction dynamics of the three individuals, as well as the solutions provided by other agents, it becomes clear that Amelia and Lily should be paired for Organizing Event Details due to their complementary strengths, meticulousness, and positive working relationship. ... This allocation allows for the efficient utilization of each person's strengths while minimizing potential conflicts and disruptions to the team's dynamics.
		Answer	Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
	Round 2 Agent 2	Reason	After considering the characteristics, strengths, and weaknesses of each individual, as well as the solutions provided by other agents, it becomes clear that the most efficient allocation of tasks would be to utilize Lily's strong network and sales experience for Seeking Sponsors and Donations, and pair Amelia and George for Organizing Event Details. ... Therefore, the most suitable allocation would be to have Lily focus on Seeking Sponsors and Donations, where her skills can be maximized without causing disruptions, and have Amelia and George work together on Organizing Event Details, with close supervision to mitigate any potential issues caused by George's behavior.
		Answer	Seeking Sponsors and Donations: Lily, Organizing Event Details: Amelia and George
	Round 2 Agent 3	Reason	After considering the characteristics, strengths, and weaknesses of each individual, as well as the solutions provided by other agents, it becomes clear that the most efficient allocation of tasks would be to utilize Amelia and Lily's collaborative and organized approach for Organizing Event Details. ... This allocation allows each individual to contribute their strengths to the tasks, minimizing the impact of their weaknesses and ensuring the success of the event.
		Answer	Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
	Final Answer		Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
DOWN	Round 1 Agent 1	Reason	The decision is based on the unique skills and strengths of each team member. Amelia's experience in attending charity galas and her ability to make connections, as well as her attention to detail, make her a strong candidate for Seeking Sponsors and Donations. ... Given their positive working relationship and respect for each other's opinions, pairing Amelia and Lily for Organizing Event Details would be an effective way to utilize their skills and minimize potential conflicts.
		Answer	Seeking Sponsors and Donations: Lily, Organizing Event Details: Amelia and George
		Confidence Score	0.89

Method	Agent	Category	Content
	Round 1 Agent 2	Reason	To efficiently accomplish both tasks, it's crucial to allocate individuals based on their strengths and working dynamics. Amelia's experience with charity events, connections, and organizational skills make her ideal for Seeking Sponsors and Donations. ... By placing George with Lily, his enthusiasm can be channeled constructively under her detail-oriented supervision, minimizing the impact of his weaknesses.
		Answer	Seeking Sponsors and Donations: Amelia, Organizing Event Details: George and Lily
		Confidence Score	0.84
	Round 1 Agent 3	Reason	The decision is based on the unique strengths and weaknesses of each team member. Amelia's experience in attending charity galas and her ability to create detailed schedules make her a strong candidate for Organizing Event Details. ... George's passion for fundraising and experience in managing sponsorships make him a good candidate for Seeking Sponsors and Donations, but his tendency to disagree with others and forget tasks make it important to pair him with a task that has fewer collaborative elements.
		Answer	Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
		Confidence Score	0.92
	Round 2 Agent 1	Reason	After considering the strengths, weaknesses, and dynamics of each team member, as well as the solutions provided by other agents, it becomes clear that the most efficient allocation of tasks would leverage the unique skills of each individual while minimizing potential conflicts. ... Despite his challenges with task management and interpersonal conflicts, his strengths can be effectively utilized in a role that plays to his advantages.
		Answer	Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
		Confidence Score	0.91
	Round 2 Agent 2	Reason	After considering the solutions from other agents and re-evaluating the strengths and weaknesses of each team member, it's clear that Amelia's skills in attending charity galas and creating detailed schedules make her a strong fit for Organizing Event Details. ...Considering these points and the fact that Lily has a strong ability to seek sponsors and donations due to her sales experience and network, and Amelia and Lily have a positive working relationship, the most efficient allocation would be to have Amelia and Lily handle Organizing Event Details, utilizing their organizational skills and positive dynamic, and George, with his passion for fundraising, focus on Seeking Sponsors and Donations, where his strengths can be maximized with minimal disruption.

Method	Agent	Category	Content
		Answer	Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
		Confidence Score	0.89
	Round 2 Agent 3	Reason	After re-evaluating the strengths, weaknesses, and dynamics of the team members, and considering the solutions from other agents, it becomes clear that Amelia's skills in attending charity galas, making connections, and attention to detail are highly valuable for Seeking Sponsors and Donations. ...Therefore, the most efficient allocation is to have Amelia focus on Seeking Sponsors and Donations, where her strengths can be fully leveraged, and to pair George with Lily for Organizing Event Details, where Lily's attention to detail and event management skills can complement George's passion for fundraising and help manage his weaknesses.
		Answer	Seeking Sponsors and Donations: Amelia, Organizing Event Details: George and Lily
		Confidence Score	0.90
	Final Answer-Vote		Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily
	Final Answer-Judge		Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily. Reason: After carefully evaluating the strengths, weaknesses, and dynamics of each team member, as well as considering the solutions provided by other agents, it is clear that the most efficient allocation of tasks would leverage the unique skills of each individual while minimizing potential conflicts. ... Considering these points, the allocation that best utilizes the strengths of each team member while minimizing disruptions is to have Amelia and Lily handle Organizing Event Details, and George focus on Seeking Sponsors and Donations, where his strengths can be maximized with minimal disruption.

Table 6: Qualitative analysis of the StrategyQA dataset

Method	Agent	Category	Content
Query	Given the story, how would you uniquely allocate each person to make sure both tasks are accomplished efficiently? Choices:‘Room Service: Vanessa, Front Desk: Angela and Miguel’, ‘Room Service: Angela, Front Desk: Miguel and Vanessa’, ‘Room Service: Miguel, Front Desk: Angela and Vanessa’ In the vibrant heart of the city, a boutique hotel thrived, its pulse maintained by the constant hum of hospitality. As the manager, my responsibility was to ensure the seamless operation of two critical areas - Room Service and Front Desk. The task of assigning these duties fell to me, with the diligent trio of Miguel, Angela, and Vanessa under my supervision. Angela, with her seven-year-long tenure in various customer service roles, harbored a graceful calm when managing guest complaints and queries. ... But Angela was not without her quirks. ... At times, Vanessa had to step in to accomplish the requests related to room service in a quick and thorough manner to compensate. Angela often found her co-worker’s unbordered enthusiasm overwhelming, especially when Vanessa would unwittingly interject her conversations with guests. Miguel, on the other hand, was the silver lining of good-natured humor and friendliness that added a personal touch to the hotel’s room service deliveries. His go-with-the-flow demeanor, however, sometimes came off as disorganized, especially to Angela who appreciated structure. ... Despite his charm and gregariousness, Miguel often had a hard time deciphering guests’ queries and concerns. This sometimes frustrated Vanessa who felt that Miguel did not accord her work ethic the respect it deserved, often leaving tasks incomplete for her to handle. Miguel’s unfamiliarity with the front desk amplified her concern, as it had never been part of his experience before. Vanessa had perfected her role at a high-end hotel reception over a span of five meticulous years. ... Vanessa meticulously paid attention to every minor detail while performing room service tasks, sometimes stepping on Angela’s toes without realizing it. To Vanessa, Angela often seemed slow, not quite pulling her weight to maintain efficiency. With these intricacies at play, as the hotel manager, I had to make conscious decisions assigning my team between Room Service and Front Desk duties. Balancing Angela, Miguel and Vanessa’s distinctive skills, patience, and perspectives, the heart of hospitality continued to beat, painting vivid memories in the hotel guest’s minds.		
Answer	Room Service: Miguel, Front Desk: Angela and Vanessa		

Method	Agent	Category	Content
MAD	Affirmative Debater	Reason	Considering the strengths and weaknesses of each person, Angela's experience in customer service and her calm demeanor make her a good fit for the Front Desk. ... Vanessa's attention to detail and professionalism make her a good fit for Room Service as well, where she can ensure that tasks are completed efficiently.
		Answer	Room Service: Miguel, Front Desk: Angela and Vanessa
	Negative Debater	Reason	I disagree with your answer. My allocation would be: Room Service: Vanessa, Front Desk: Angela and Miguel Here's my reasoning: * Vanessa's attention to detail and professionalism make her a perfect fit for Room Service, where she can ensure that tasks are completed efficiently and accurately. ... * Angela's experience in customer service and her calm demeanor make her a good fit for the Front Desk, where she can handle guest complaints and queries in a professional and respectful manner. ... * Miguel's charm and friendliness are also well-suited for the Front Desk, where he can interact with guests and provide a warm welcome. ... Overall, ... Vanessa's efficiency and attention to detail in Room Service can help to compensate for Angela's occasional misplacement of minor notes, while Angela and Miguel's combined strengths at the Front Desk can provide a warm and professional welcome for guests.
		Answer	Room Service: Vanessa, Front Desk: Angela and Miguel
	Moderator	Reason	The negative side's allocation takes into account the strengths and weaknesses of each team member more effectively. ... This allocation also addresses the potential issues with each team member, such as Angela's occasional misplacement of minor notes and Miguel's difficulty with deciphering guests' queries and concerns.
		Answer	Room Service: Vanessa, Front Desk: Angela and Miguel
	Final Answer		Room Service: Vanessa, Front Desk: Angela and Miguel
Debate	Round 1 Agent 1	Reason	To allocate each person efficiently, we need to consider their strengths and weaknesses. ... Miguel's charm and friendliness are best utilized in room service, where he can interact with guests in a more relaxed environment and work at his own pace without the strict structure of the front desk.
		Answer	Room Service: Miguel, Front Desk: Angela and Vanessa

Method	Agent	Category	Content
	Round 1 Agent 2	Reason	To efficiently allocate tasks, we need to consider the strengths and weaknesses of each person. ... Therefore, the most efficient allocation would be to assign Vanessa to Room Service, where her attention to detail and professionalism can shine, and Angela and Miguel to the Front Desk, where Angela's customer service skills can be utilized and Miguel's charm can be an asset.
		Answer	Room Service: Vanessa, Front Desk: Angela and Miguel
	Round 1 Agent 3	Reason	To allocate each person efficiently, we need to consider their strengths and weaknesses. ... However, considering the need for balance and the fact that Miguel might struggle with the structure of the Front Desk, it's better to pair Angela with Miguel at the Front Desk, where Angela's organizational skills can help Miguel, and Vanessa can handle Room Service, utilizing her strengths without overwhelming others.
		Answer	Room Service: Vanessa, Front Desk: Angela and Miguel
	Round 2 Agent 1	Reason	Considering the strengths and weaknesses of each person, Angela's customer service skills are best utilized at the Front Desk, but her disorganization could be mitigated by pairing her with a detail-oriented person. ... Given these considerations and the analysis from other agents, it's clear that Angela and Miguel's skills complement each other at the Front Desk, with Angela handling customer interactions and Miguel providing charm, while Vanessa's strengths are best utilized in Room Service, ensuring efficiency and attention to detail.
		Answer	Room Service: Vanessa, Front Desk: Angela and Miguel
	Round 2 Agent 2	Reason	Considering the strengths and weaknesses of each person, Angela's customer service skills make her a strong fit for the Front Desk, while Miguel's charm can be leveraged in Room Service. ... The most efficient allocation, considering all perspectives, would be to assign Vanessa to Room Service, where her attention to detail can ensure tasks are completed efficiently, and Angela and Miguel to the Front Desk, where Angela's customer service skills and organizational abilities can help mitigate Miguel's weaknesses.
		Answer	Room Service: Vanessa, Front Desk: Angela and Miguel
	Round 2 Agent 3	Reason	Considering the strengths and weaknesses of each person, Angela's calm demeanor and customer service skills make her a strong fit for the Front Desk. ... Assigning Vanessa to Room Service leverages her strengths, and pairing Angela with Miguel at the Front Desk utilizes Angela's customer service skills and Miguel's charm, with Angela's organizational skills potentially helping to mitigate Miguel's disorganization.
		Answer	Room Service: Vanessa, Front Desk: Angela and Miguel
	Final Answer		Room Service: Vanessa, Front Desk: Angela and Miguel

Method	**Agent**	**Category**	**Content**
DOWN	Round 1 Agent 1	Reason	To efficiently allocate tasks, it's essential to consider the strengths and weaknesses of each team member. Angela excels in customer service, particularly at the Front Desk, due to her calm demeanor and respectful tone. ... Additionally, her efficiency and thoroughness in room service tasks, although sometimes stepping on Angela's toes, demonstrate her capability to handle detailed work.
		Answer	Room Service: Miguel, Front Desk: Angela and Vanessa
		Confidence Score	0.91
	Final Answer		Room Service: Miguel, Front Desk: Angela and Vanessa

Table 7: Qualitative analysis of the MUSR dataset