

# 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 debate framework that selectively activates debate based on the confidence score of the agent’s initial response. Debate is activated only for queries requiring further deliberation, during which agents refine their outputs by referencing peer responses and 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 unnecessary 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

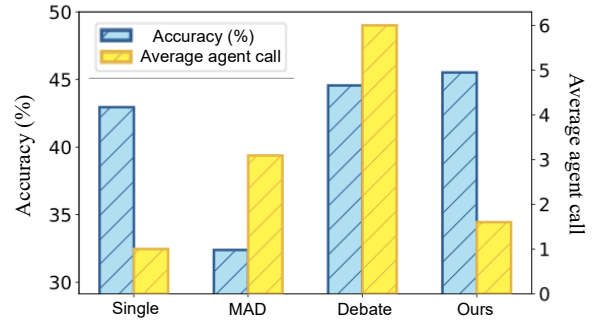


Figure 1: Comparison of accuracy and average agent calls across various multiagent debate methods

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).

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 or unnecessary 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 De-

bate Only When Necessary (DOWN), an adaptive multiagent collaboration framework that identifies queries requiring debate and selectively facilitates the debate process accordingly. This framework employs the confidence score to quantify the internal certainty of LLM outputs, a measure that has been extensively utilized to enhance model performance and trustworthiness (Razghandi et al., 2025; Taubenfeld et al., 2025; Chen et al., 2024a). In the DOWN framework, the initial agent response is generated as the first step. A confidence score is computed during response generation and used to determine whether debate should be initiated. If the confidence score exceeds a threshold, the debate is skipped, and we regard the initial response as a final answer. If further refinement is required, agents engage in debate to enhance response accuracy. When the collaboration begins, a confidence-guided debate is conducted, during which agents refine their responses by considering both the responses of other participating agents and their confidence scores. This approach encourages the refinement of responses by utilizing the most persuasive aspects of agent responses.

Our experiments employ models of varying scales, including approximately 8B and 70B parameter models, as well as GPT-4o-mini, evaluated on the MUSR (Sprague et al., 2024) and StrategyQA (Geva et al., 2021) benchmarks. The results demonstrate that adaptive debate invoking significantly reduces 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. Our contribution is threefold:

- We propose Debate Only When Necessary (DOWN), an adaptive multiagent framework that selectively initiates debate based on the initial response. To the best of our knowledge, this is the first study to explore conditional debate in multiagent systems grounded in initial responses, maximizing efficiency while preserving the benefits of discussion.
- Extensive experiments across diverse model sizes and configurations establish the effectiveness of the mechanism. We further find that adaptively engaging debate contributes to mitigating error propagation.
- We establish that the confidence-guided de-

bate process enables the selective integration of reliable responses, emphasizing the effectiveness of multiagent collaboration.

## 2 Related Work

**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.

**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) introduce a shortcut mechanism that shares similarities with our approach. However, both methods still incur non-trivial agent calls. The former requires neighbor interactions, and the latter relies on self-consistency, with both involving 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.

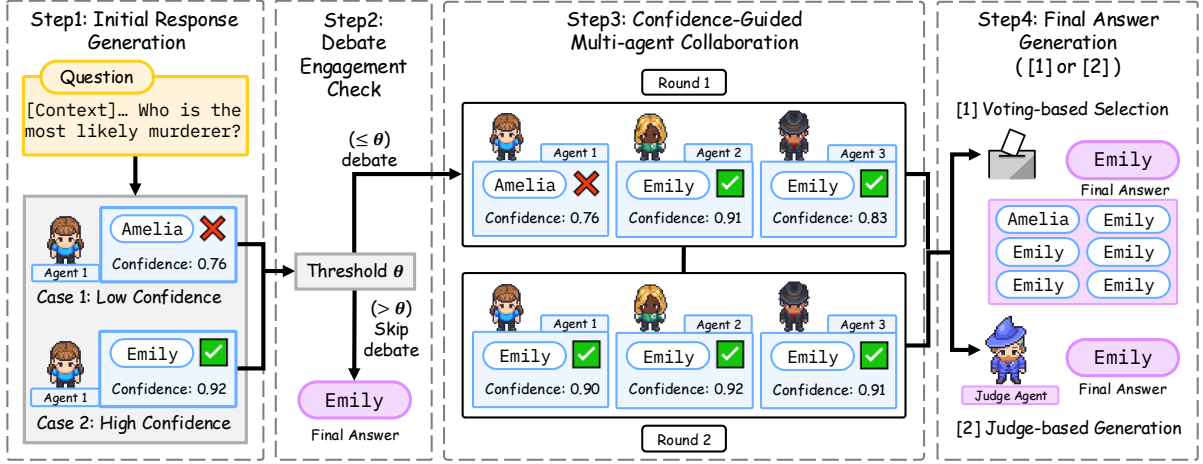


Figure 2: Overview of the Debate Only When Necessary (DOWN) framework. DOWN consists of four stages: (1) the initial agent generates a response, during which the model’s confidence score is extracted. (2) if the confidence score exceeds a threshold value, the response is accepted without debate to improve efficiency, otherwise a multiagent debate is activated. (3) agents refine their responses by referencing peer outputs and associated confidence scores. (4) the final answer is selected via majority voting or designated judge agent.

### 3 Debate Only When Necessary

The collaboration framework consists of four steps: (1) initial response generation, (2) debate engagement check, (3) confidence-guided multiagent collaboration, and (4) final response generation. The confidence score derived from the model during initial response generation determines its subsequent progression. If the score exceeds a predefined threshold, the further debate process is skipped. Otherwise, we proceed with the debate, allowing agents to engage in confidence-guided debate with other agents. We illustrate the overall multiagent collaboration framework in Figure 2.

#### 3.1 System Pipeline

**Step 1: Initial Response Generation** The model’s confidence score obtained during initial response generation determines the progression of the collaboration process. Based on its importance, we adopt two strategies for multiagent configurations: (1) A homogeneous-agent configuration where all collaborating agents use the same model (e.g., Llama) and (2) a mixed-agent configuration where the initial agent is randomly selected for each query (e.g., Llama, Qwen, GPT-4o-mini).

For a given query  $q$ , an initial agent  $\mathcal{A}_1$  from the set of agents  $\{\mathcal{A}\}_{i=1}^N$  generates an initial response  $r_1$  with the reason for the answer. During this process, a confidence score  $c_1$  is extracted to quantify the model’s certainty. We first obtain the token logit  $L(t_i)$ , where  $t_i \in r_1$ , from the hidden representa-

tion 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:

$$c_1 = \frac{1}{|r_1|} \sum_{i=1}^{|r_1|} P(t_i), \quad \text{where } P(t_i) = \frac{e^{L(t_i)}}{\sum_j e^{L(t_j)}}. \quad (1)$$

For models that cannot access internal logits, we utilize verbalized confidence as an alternative method.

**Step 2: Debate Engagement Check** Along with the initial response, we extract a confidence score  $c_1$  reflecting the model’s certainty in its answer. The confidence score is compared against a threshold score  $\theta$ :

- (1) High confidence ( $> \theta$ ): The response is accepted as the final response without further collaboration, optimizing efficiency by avoiding unnecessary computations.
- (2) Low confidence ( $\leq \theta$ ): The initial response is deemed unreliable, activating the collaboration process to improve response quality.

In this context, the threshold is a hyperparameter, where a higher threshold places more emphasis on accuracy, while a lower threshold prioritizes efficiency.

### Step 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 generate their responses in the first round, the second round begins. Each agent is given responses from the other two agents, excluding its own. Using this additional context, each agent refines its reasoning and generates an updated response. This allows each agent to leverage confident and compelling arguments in the updated response.

**Step 4: Final Answer Generation** The responses generated in Step 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 Threshold Selection Strategy

To select the optimal confidence threshold  $\theta^* \in \Theta$  that jointly maximizes predictive accuracy and inference-time efficiency, we introduce a scoring function that incorporates two soft penalty terms for both objectives. To support high-performance threshold selection, we define a statistically grounded high-performance region for DOWN, derived by computing the one-sided 95% Wilson lower bound of the maximum observed accuracy. Threshold candidates whose accuracy lies within this region are considered statistically indistinguishable from the best-performing value, whereas those outside the region are penalized via a soft penalty term. Let  $A_\theta = \frac{k}{N}$  and  $S_\theta = \frac{m}{N}$  denote the accu-

racy and debate skip rate (i.e. efficiency) for a given threshold value  $\theta$  in the candidate set  $\Theta$ . Given the highest observed accuracy  $A_\theta^{\max}$  obtained from  $k^{\max}$  correct predictions, the Wilson lower bound  $\tilde{A}_\theta^{\max}$  is computed as:

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

where  $z = \Phi^{-1}(0.95) = 1.645$ . Thresholds with raw accuracy below  $\tilde{A}_\theta^{\max}$  are subject to soft penalization.

In parallel, to ensure a balanced trade-off between computational efficiency and performance improvement through debate, we penalize thresholds whose skip rate  $S_\theta$  falls outside an interval  $[s_{\min}, 1 - s_{\min}]$ . Here,  $s_{\min}$  denotes the proportion symmetrically excluded from the lower and upper bounds of the unit interval  $[0, 1]$ . The total penalty term is defined as:  $P_\theta = \max(0, \tilde{A}_\theta^{\max} - A_\theta) + \max(0, s_{\min} - S_\theta) + \max(0, S_\theta - (1 - s_{\min}))$ .

With both penalty terms defined, we compute the overall utility score for each threshold candidate. To ensure that accuracy and efficiency contribute comparably to the final score, we normalize both metrics via min-max scaling:

$$\hat{A}_\theta = \frac{A_\theta - A_\theta^{\min}}{A_\theta^{\max} - A_\theta^{\min} + \epsilon}, \quad \hat{S}_\theta = \frac{S_\theta - S_\theta^{\min}}{S_\theta^{\max} - S_\theta^{\min} + \epsilon}, \quad (3)$$

where  $\epsilon \ll 1$  is a small constant added for numerical stability. We compute the final score as the sum of the normalized accuracy  $\hat{A}_\theta$  and efficiency  $\hat{S}_\theta$ , with a penalty term  $P_\theta$  scaled by a weight factor  $\lambda$ . The optimal threshold is then determined by maximizing this score:  $\theta^* = \arg \max_{\theta \in \Theta} (\hat{A}_\theta + \hat{S}_\theta - \lambda \cdot P_\theta)$ .

## 4 Experiments

### 4.1 Experimental Setup

**Models.** We evaluate two agent configurations: homogeneous and mixed. The homogeneous setup leverages a single model across all agents, experimenting with Llama-3.1 8B (meta-llama/Llama-3.1-8B-Instruct), Ministral 8B (mistralai/Ministral-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). The mixed configuration utilizes Qwen-2.5 72B, Llama-3.3 70B, and GPT-4o-mini,



| Method           | Llama-3.1 8B |      | Ministral 8B |      | GPT-4o-mini  |      | Llama-3.3 70B |      | Qwen-2.5 72B |      |
|------------------|--------------|------|--------------|------|--------------|------|---------------|------|--------------|------|
|                  | Acc.         | AC   | Acc.         | AC   | Acc.         | AC   | Acc.          | AC   | Acc.         | AC   |
| Single-CoT       | 42.95        | 1.00 | 51.06        | 1.00 | 55.75        | 1.00 | 56.33         | 1.00 | 57.80        | 1.00 |
| Self-refine      | 39.46        | 6.00 | 36.90        | 6.00 | 54.29        | 6.00 | 53.67         | 6.00 | 58.47        | 6.00 |
| Self-Consistency | 44.70        | 9.00 | 48.24        | 9.00 | 55.88        | 9.00 | <b>58.18</b>  | 9.00 | 58.29        | 9.00 |
| MAD              | 32.39        | 3.09 | 28.67        | 3.01 | 43.23        | 3.02 | 51.22         | 3.00 | 49.13        | 3.04 |
| Debate           | 44.56        | 6.00 | 48.54        | 6.00 | <u>57.32</u> | 6.00 | 57.28         | 6.00 | 58.69        | 6.00 |
| DOWN-Vote        | <u>45.51</u> | 1.50 | <b>53.71</b> | 1.48 | 57.09        | 1.80 | <u>57.80</u>  | 1.02 | <u>59.39</u> | 1.28 |
| DOWN-Judge       | <b>45.52</b> | 1.60 | <b>53.71</b> | 1.57 | <b>57.35</b> | 1.96 | <u>57.80</u>  | 1.03 | <b>59.52</b> | 1.34 |

Table 1: Comparison of accuracy (Acc.) and average agent calls (AC) across single-agent methods, multiagent debate systems, and our proposed approach on the MUSR dataset. Single CoT, Self-refine, and Self-consistency are single model-based approaches, while MAD, Debate, and DOWN are multiagent debate-based systems.

| Method           | Llama-3.1 8B |      | Ministral 8B |      | GPT-4o-mini  |      | Llama-3.3 70B |      | Qwen-2.5 72B |      |
|------------------|--------------|------|--------------|------|--------------|------|---------------|------|--------------|------|
|                  | Acc.         | AC   | Acc.         | AC   | Acc.         | AC   | Acc.          | AC   | Acc.         | AC   |
| Single-CoT       | <u>70.74</u> | 1.00 | 67.69        | 1.00 | 78.17        | 1.00 | 80.35         | 1.00 | <u>78.40</u> | 1.00 |
| Self-refine      | 69.54        | 6.00 | 67.69        | 6.00 | 76.42        | 6.00 | 77.73         | 6.00 | 78.17        | 6.00 |
| Self-Consistency | 68.56        | 9.00 | <u>68.12</u> | 9.00 | 79.48        | 9.00 | <u>80.79</u>  | 9.00 | 77.29        | 9.00 |
| MAD              | 44.54        | 4.66 | 57.64        | 3.73 | 70.31        | 3.38 | 79.04         | 3.07 | 73.80        | 3.24 |
| Debate           | 70.08        | 6.00 | <b>70.74</b> | 6.00 | 79.04        | 6.00 | 80.35         | 6.00 | <b>79.91</b> | 6.00 |
| DOWN-Vote        | <b>71.18</b> | 2.53 | <u>68.12</u> | 3.16 | <b>80.79</b> | 1.92 | <b>82.53</b>  | 1.07 | 77.73        | 2.64 |
| DOWN-Judge       | 69.87        | 2.83 | <u>68.12</u> | 3.59 | <u>79.91</u> | 2.10 | <b>82.53</b>  | 1.08 | 77.73        | 2.97 |

Table 2: Comparison of accuracy (Acc.) and average agent calls (AC) across single-agent methods, multiagent debate systems, and our proposed approach on the StrategyQA dataset

with the order of model selection randomized for each query. Each debate round consists of three agents, and we conduct a two-round debate where responses are generated in the first round and refined in the second.

**Tasks.** For evaluation, we utilize benchmarks specifically 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.

**Baselines.** We compare our approach against multiple established reasoning frameworks, along with a single LLM CoT (Kojima et al., 2022) baseline. (1) **Self-refine** (Madaan et al., 2023): Generating an initial response and iteratively refining its output through self-feedback. We design an environment composed of two rounds,

with each round comprising three agents. (2) **Self-Consistency** (Wang et al., 2023): Sampling diverse reasoning paths and deriving the final answer by selecting the most consistent one through majority voting. We sample nine responses for each query. (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): Engaging agents in a structured debate, iteratively refining their responses by incorporating insights from previous exchanges. We design an environment composed of two rounds, with each round comprising three agents.

## 4.2 Implementation Details

We set the temperature to 0.0 to ensure deterministic response generation and limit the maximum sequence length to 512 tokens. The experimental setup maintains consistent configurations across all models. For the  $\theta^*$  selection, we define the candidate set as  $\Theta = \{0.7, 0.8, 0.9\}$ , starting from a threshold that yields a debate skip rate  $S_\theta$  close to 1. The symmetric margin  $s_{\min}$  and the penalty scaling factor  $\lambda$  are set to 0.1 and 15. For the evaluation metrics, we use accuracy as the metric, while

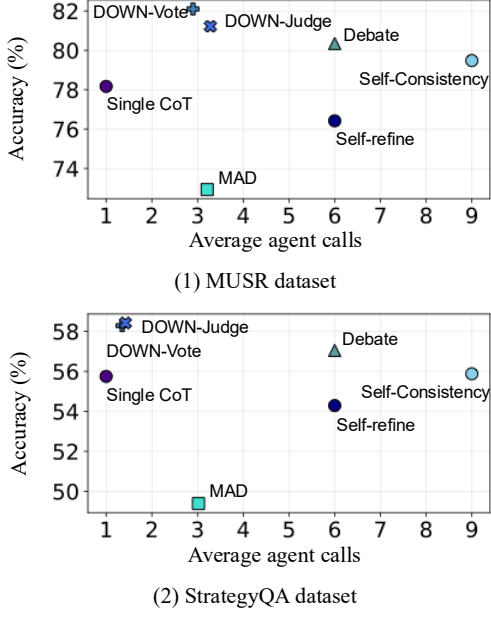


Figure 3: Comparison of multiagent debate system performance in a mixed-model configuration. The configuration includes Llama3.3-70B, Qwen-2.5 72B, and GPT-4o-mini, with the model order randomized for each query. For single model-based approaches, we present the results of GPT-4o-mini.

efficiency is assessed based on the average number of agent calls or debate skip rate. The experiments are conducted using four 48GB A6000 GPUs. All prompts used in the experiments are provided in Table 6.

## 5 Results and Analysis

### 5.1 Main Results

**Homogeneous-Model Configuration** Table 1 presents the results on the MUSR dataset. Unlike prior approaches that require three to nine agent calls per query, our framework achieves comparable or superior performance with only 1.4 calls on average, representing over a sixfold improvement in efficiency. The results in Table 2 on the StrategyQA dataset further support the effectiveness of our approach. Compared to Debate and Self-Consistency, our approach achieves comparable or superior performance while significantly reducing the average number of agent calls. Moreover, our framework demonstrates strong robustness across different model families and parameter sizes, confirming its applicability across a wide range of architectures. These findings highlight that selectively invoking debate, rather than applying it uniformly or relying solely on single agent responses,

| LLM Agent     | 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 changes before and after debate across multiagent collaboration methods. We denote a correct answer by ✓ and an incorrect answer by ✗.

offers a more principled balance between accuracy and efficiency. Interestingly, MAD exhibits lower accuracy compared to other debate-based methods. 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.

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

**Comparison of Final Answer Generation Strategies** Voting-based selection and judge-based evaluation are two strategies for consolidating multiagent 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

| Model         | Method | Original Debate | Conditional Debate | Skip Rate |
|---------------|--------|-----------------|--------------------|-----------|
| Llama-3.3 70B | MAD    | 79.04           | <b>79.91</b>       | 59.83%    |
|               | Debate | 80.35           | <b>83.41</b>       | 68.56%    |
| Qwen-2.5 72B  | MAD    | 73.80           | <b>76.86</b>       | 51.53%    |
|               | Debate | <b>79.91</b>    | 79.48              | 45.85%    |

Table 4: Results of applying conditional debate to existing debate systems on the StrategyQA dataset

it achieves comparable accuracy with fewer computational resources.

## 5.2 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 confidence threshold 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-3.3-70B model, 87.43% of changed answers reflect successful corrections of initially incorrect predictions. In contrast, MAD shows a high rate of incorrect changes, frequently revising correct initial answers into incorrect ones, consistent with the trends observed in § 5.1. The Debate baseline shows a similar rate of correct and incorrect changes, indicating a limited capacity to prioritize reliable peer inputs. In contrast, the high success rate of corrections achieved by DOWN stems from two core design principles. It initiates debate only when the model’s confidence is low, thereby avoiding unnecessary changes to already correct answers. When debate is triggered, it refers to peer responses with their associated confidence scores, enabling the model to incorporate more reliable inputs.

## 5.3 Effects of Conditional Debate in Multiagent Debates

To assess the effectiveness of conditional debate, we apply it to MAD and Debate frameworks and evaluate performance on the StrategyQA dataset using models with approximately 70B parameters. Our findings in Table 4 reveal that applying conditional debate to existing debate methods mostly improves performance. In particular, the Debate framework achieves about a 3% point accuracy increase on LLaMA-3.3 70B, while MAD exhibits a similar improvement on Qwen-2.5 72B. Despite these gains, debate skip rates remain high at 59.83%

| Method  | Acc.  | AC   |
|---|-------|------|
| DOWN ( $\theta^*$ )                               | 71.18 | 2.53 |
| <i>Debate</i>                                     |       |      |
| (1) w/o confidence-based scoring in debate        | 68.12 | 2.46 |
| (2) w/o multi-agent debate (single response only) | 70.74 | 1.00 |
| (3) w/o debate skipping (debate only)             | 70.74 | 6.00 |
| <i>Threshold</i>                                  |       |      |
| (4) w/ lowered threshold ( $\theta^* - 0.1$ )     | 69.87 | 1.20 |
| (5) w/ raised threshold ( $\theta^* + 0.1$ )      | 71.62 | 3.53 |

Table 5: Ablation study on confidence score using the StrategyQA dataset

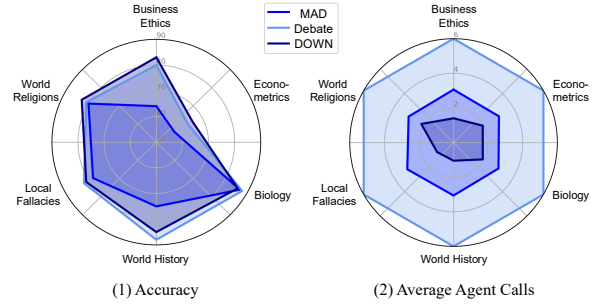


Figure 4: Accuracy and average agent calls (AC) of multiagent debate methods across six MMLU domains

and 51.53%, respectively, demonstrating remarkable efficiency improvements. These results demonstrate that conditional debate not only enhances efficiency but also positively impacts overall model performance by selectively engaging discussions at appropriate points.

## 5.4 Ablation Study

To analyze the contribution of each component in the DOWN framework, we conduct an ablation study on the StrategyQA using the LLaMA-3.1 8B model. The results are shown in Table 5. We examine the role of confidence scores shared among agents during debate. Removing these signals leads to a 3.06% point drop in accuracy, indicating that confidence serves as an informative cue for selectively incorporating peer responses during the debate process. When the debate process is either entirely removed or enforced for every input, the accuracy drops to 70.74 in both cases. These findings indicate that always engaging in debate is redundant, yielding no improvement over the single agent baseline. The next analysis focuses on the threshold selection strategy. To assess its effectiveness, we vary the threshold by  $\pm 0.1$  and observe the resulting impact on performance. Modifying the threshold above or below the selected value reveals a trade-off between accuracy and average

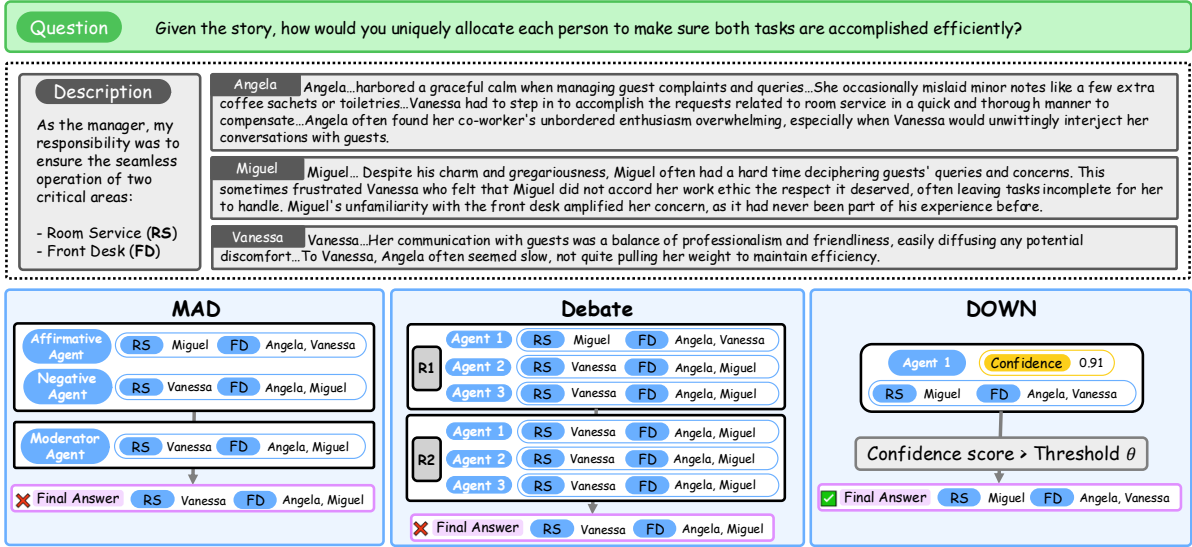


Figure 5: Qualitative analysis of the MUSR dataset

agent calls. Specifically, the scores obtained for the threshold values  $\{\theta^* - 0.1, \theta^*, \theta^* + 0.1\}$  yields  $\{0.29, 1.19, 1.00\}$ , respectively. The chosen threshold  $\theta^*$  yields the highest overall score, effectively balancing accuracy and efficiency. It maintains accuracy within a statistically reliable range, as determined by the Wilson lower bound, while substantially reducing computational cost.

### 5.5 Generalization analysis on MMLU

To assess the generalization ability of the DOWN framework across diverse task types and knowledge domains, we additionally evaluate its performance on MMLU using GPT-4o-mini. The performance of different multiagent debate systems across six domains is presented in Figure 4. Experimental results show that DOWN achieves performance comparable to that of the Debate method across tasks. Given that the Debate approach consistently 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. This highlights its potential to serve as a scalable and effective alternative to full multiagent debate systems.

### 5.6 Qualitative Analysis

Figure 5 presents a qualitative comparison of responses generated by different debate systems on the MUSR dataset. The results show that existing debate mechanisms introduce unnecessary modifi-

cations. 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 selectively engaging in debate under high-confidence conditions prevents unnecessary modifications while maintaining efficiency. To further substantiate these findings, we present additional qualitative analysis in Appendix B.

## 6 Conclusion

This work proposed the DOWN framework to address the computational inefficiencies and error propagation challenges in multiagent collaboration. By conditionally activating debate based on the model confidence score, the approach significantly reduced computational overhead while preserving or even improving performance. The results showed that conditional debate enhanced efficiency and mitigated cascading errors, leading to more stable reasoning behavior. Furthermore, the confidence-guided multiagent debate amplified the influence of reliable responses on final decisions. These findings established DOWN as an effective optimization strategy, offering a high-performance and efficient solution for multiagent collaboration systems.



## Limitations

While our proposed DOWN framework demonstrates strong efficiency and robustness, several limitations remain. 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.
- 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.
- Huaoli, 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 Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder,

|     |   |     |
|-----|---|-----|
| 669 | Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. <a href="#">Self-refine: Iterative refinement with self-feedback</a> . In <i>Advances in Neural Information Processing Systems</i> , volume 36, pages 46534–46594. Curran Associates, Inc.   | 722 |
| 670 |   | 723 |
| 671 |   | 724 |
| 672 |   | 725 |
| 673 |   | 726 |
| 674 | Marvin Minsky. 1988. <i>Society of mind</i> . Simon and Schuster.   | 727 |
| 675 |   |     |
| 676 | Ali Razghandi, Seyed Mohammad Hadi Hosseini, and Mahdieh Soleymani Baghshah. 2025. Cer: Confidence enhanced reasoning in llms. <i>arXiv preprint arXiv:2502.14634</i> .   | 728 |
| 677 |   | 729 |
| 678 |   | 730 |
| 679 |   | 731 |
| 680 | Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. <i>arXiv preprint arXiv:2408.03314</i> .  | 732 |
| 681 |   | 733 |
| 682 |   | 734 |
| 683 |   |     |
| 684 | Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. <a href="#">MuSR: Testing the limits of chain-of-thought with multistep soft reasoning</a> . In <i>The Twelfth International Conference on Learning Representations</i> .  | 735 |
| 685 |   | 736 |
| 686 |   | 737 |
| 687 |   | 738 |
| 688 |   | 739 |
| 689 | 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. <i>arXiv preprint arXiv:2410.09403</i> .   | 740 |
| 690 |   | 741 |
| 691 |   | 742 |
| 692 |   | 743 |
| 693 |   | 744 |
| 694 |   | 745 |
| 695 | Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. <i>arXiv preprint arXiv:2502.06233</i> .   | 746 |
| 696 |   | 747 |
| 697 |   |     |
| 698 |   |     |
| 699 | 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. <i>arXiv preprint arXiv:2501.06322</i> .  | 748 |
| 700 |   | 749 |
| 701 |   | 750 |
| 702 |   | 751 |
| 703 |   | 752 |
| 704 | Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. <a href="#">Mixture-of-agents enhances large language model capabilities</a> . In <i>The Thirteenth International Conference on Learning Representations</i> .  | 753 |
| 705 |   | 754 |
| 706 |   |     |
| 707 |   |     |
| 708 |   |     |
| 709 | Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. <a href="#">Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?</a> In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.  | 755 |
| 710 |   | 756 |
| 711 |   | 757 |
| 712 |   | 758 |
| 713 |   | 759 |
| 714 |   | 760 |
| 715 |   | 761 |
| 716 | Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. <a href="#">Self-consistency improves chain of thought reasoning in language models</a> . In <i>The Eleventh International Conference on Learning Representations</i> .  | 762 |
| 717 |   | 763 |
| 718 |   | 764 |
| 719 |   | 765 |
| 720 |   | 766 |
| 721 |   |     |
|     | Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Auto-gen: Enabling next-gen llm applications via multi-agent conversation framework. <i>arXiv preprint arXiv:2308.08155</i> .   | 767 |
|     |   | 768 |
|     | Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. <a href="#">Examining inter-consistency of large language models collaboration: An in-depth analysis via debate</a> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 7572–7590, Singapore. Association for Computational Linguistics.   | 769 |
|     |   | 770 |
|     |   | 771 |
|     |   | 772 |
|     | Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. <a href="#">React: Synergizing reasoning and acting in language models</a> . In <i>The Eleventh International Conference on Learning Representations</i> .  | 773 |
|     |   | 774 |
|     | Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. <a href="#">Exchange-of-thought: Enhancing large language model capabilities through cross-model communication</a> . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 15135–15153, Singapore. Association for Computational Linguistics.  | 775 |
|     |   | 776 |
|     | Wei Zhou, Mohsen Mesgar, Annemarie Friedrich, and Heike Adel. 2025. <a href="#">Efficient multi-agent collaboration with tool use for online planning in complex table question answering</a> . In <i>Findings of the Association for Computational Linguistics: NAACL 2025</i> , 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. <a href="#">Mindstorms in natural language-based societies of mind</a> . In <i>R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models</i> . |     |
|     |   |     |
|     | <b>A Prompts leveraged in DOWN framework</b>  |     |
|     |   |     |
|     | We present the prompts utilized in our experiments in Table 6. These prompts are applied throughout the DOWN framework to ensure structured and consistent response generation.   |     |
|     |   |     |
|     | <b>B Qualitative Analysis of Debate Systems</b>   |     |
|     |   |     |
|     | Additional qualitative analysis for each multiagent collaboration method is presented in Table 7, Table 8, and Table 9.   |     |

| Purpose                                  | Prompt  |
|--|---|
| 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                            | 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. |

Table 6: Prompt configuration leveraged in our experiments

Table 7 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 8 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 9 follows a trend similar to Table 7. 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<br>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<br>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<br>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<br>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 7: Qualitative analysis of the StrategyQA dataset

| Method | Agent  | Category | Content  |
|--------|--|----------|--|
| Query  | <p>Given the story, how would you uniquely allocate each person to make sure both tasks are accomplished efficiently?</p> <p>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’</p> <p>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.</p> <p>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. ...</p> <p>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.</p> <p>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.</p> <p>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.</p> |          |  |
| Answer | Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily  |          |  |
| MAD    | Affirmative Debater  | Reason   | 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. |
|        |  | Answer   | Seeking Sponsors and Donations: Amelia, Organizing Event Details: George and Lily  |

| 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  |
|        | <b>Final Answer</b> |          | 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<br>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<br>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<br>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   |
|        | <b>Final Answer</b> |                  | Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily   |
| DOWN   | Round 1<br>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<br>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<br>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<br>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<br>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<br>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   |
|        |                           | <b>Final Answer-Vote</b> | Seeking Sponsors and Donations: George, Organizing Event Details: Amelia and Lily  |
|        | <b>Final Answer-Judge</b> |                          | 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 8: Qualitative analysis of the StrategyQA dataset

| Method | Agent | Category | Content   |
|--------|-------|----------|---|
| Query  |       |          | <p>Given the story, how would you uniquely allocate each person to make sure both tasks are accomplished efficiently?</p> <p>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'</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p> |
| 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:<br><br>Room Service: Vanessa, Front Desk: Angela and Miguel<br><br>Here's my reasoning:<br><br>* 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. ...<br>* 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. ...<br>* Miguel's charm and friendliness are also well-suited for the Front Desk, where he can interact with guests and provide a warm welcome. ...<br><br>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<br>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<br>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<br>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<br>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<br>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  |
|        | <b>Final Answer</b> |          | Room Service: Vanessa, Front Desk: Angela and Miguel  |

| Method | Agent               | Category         | Content   |
|--------|---------------------|------------------|---|
| DOWN   | Round 1<br>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  |
|        | <b>Final Answer</b> |                  | Room Service: Miguel, Front Desk: Angela and Vanessa  |

Table 9: Qualitative analysis of the MUSR dataset