Voting or Consensus? Decision-Making in Multi-Agent Debate

Lars Benedikt Kaesberg[1,*], **Jonas Becker**[1,2], **Jan Philip Wahle**[1], **Terry Ruas**[1], **Bela Gipp**[1]
[1]University of Göttingen, Germany; [2]LKA NRW, Germany
[*]**Correspondence:** l.kaesberg@uni-goettingen.de

Abstract

Much of the success of multi-agent debates depends on carefully choosing the right parameters. The decision-making protocol stands out as it can highly impact final model answers, depending on how decisions are reached. Systematic comparison of decision protocols is difficult because many studies alter multiple discussion parameters beyond the protocol. So far, it has been largely unknown how decision-making influences different tasks. This work systematically evaluates the impact of seven decision protocols (e.g., majority voting, unanimity consensus). We change only one variable at a time — the decision protocol — to analyze how different methods affect the collaboration between agents and measure differences in knowledge and reasoning tasks. Our results show that voting protocols improve performance by 13.2% in reasoning tasks and consensus protocols by 2.8% in knowledge tasks compared to other decision protocols. Increasing the number of agents improves performance, while more discussion rounds before voting reduce it. To improve decision-making by increasing answer diversity, we propose two new methods, All-Agents Drafting (AAD) and Collective Improvement (CI). Our methods improve task performance by up to 3.3% with AAD and up to 7.4% with CI. This work demonstrates the importance of decision-making in multi-agent debates beyond scaling.

1 Introduction

Humans are inherently social, and collaboration has been key to innovation and progress. We know that generating solutions together is only beneficial if we can effectively select, agree, and commit to them. History, sociology, and psychology have long demonstrated how different decision-making processes influence collective outcomes (Jones, 1994; List, 2022). Multi-agent systems form a parallel to human behavior by solving problems collectively via debate. However, so far, few studies have investigated how decision-making influences large language model (LLM) collaboration and their ability to problem-solve.

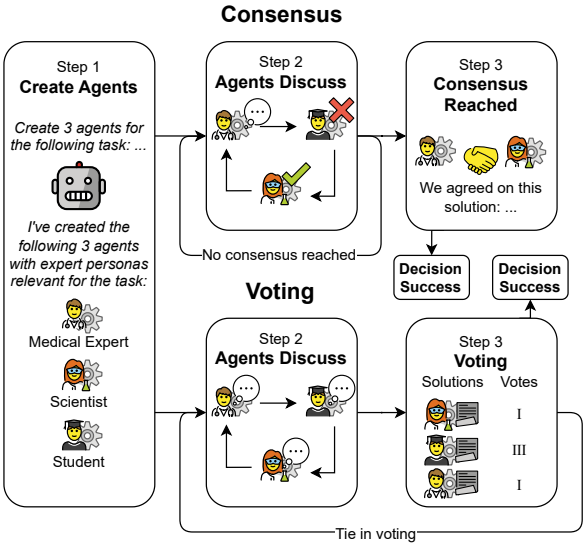

Figure 1: Illustration of voting and consensus-based decision protocols used in this study.

Current approaches often apply decision strategies like majority voting (Yang et al., 2024b) or consensus (Yin et al., 2023) indiscriminately to various tasks. Consensus strategies refine decisions while voting approaches choose from proposed solutions. Prior work often treats these strategies as fixed variables without considering problem-specific characteristics (Yin et al., 2023). We show that changes in decision protocols lead to markedly different results among tasks. Thus, we argue that decision-making is central to multi-agent processes.

This study systematically quantifies the effectiveness of decision-making protocols within multi-agent debates on knowledge tasks (i.e., MMLU (Hendrycks et al., 2021), MLLU-Pro (Wang et al., 2024b), GPQA (Rein et al., 2023)) and reasoning tasks (i.e., SQuAD 2.0 (Rajpurkar et al., 2018), StrategyQA (Geva et al., 2021), MuSR (Sprague et al., 2023)). In contrast to prior work, which

Findings of the Association for Computational Linguistics: ACL 2025, pages 11640–11671

varies the decision protocol concurrently to other parameters (Yang et al., 2024b; Yin et al., 2023), we use a consistent setup to quantify changes for which only the decision mechanism is responsible. Specifically, we explore the impact of three consensus (Chen et al., 2023) and four voting (Yang et al., 2024b) methods and highlight their task-specific strengths and weaknesses. An overview of how consensus and voting decision protocols work in our setup can be seen in Figure 1. Throughout our experiments, we observe that answer diversity and independent answer generation have a marked impact on decision-making. To reap the benefits of these factors, we propose two methods: All-Agents Drafting (AAD) and Collective Improvement (CI). AAD ensures each agent independently drafts an initial solution before any interaction, promoting distinct reasoning paths. CI structures agent collaboration through iterative refinement while preventing excessive communication to avoid bias towards similar answers between agents.

Our experiments show that consensus protocols perform better in knowledge tasks, improving performance by 2.8%, and voting protocols perform better in reasoning tasks, improving results by 13.2%. AAD increases accuracy by about 3.3%, while CI further improves effectiveness, leading to an 7.4% performance boost. Our results underline the role of decision protocols for multi-agent experiments. We recommend consensus strategies for knowledge tasks and voting for reasoning tasks while generally using AAD or CI to improve answer diversity. Our investigations into the role of decision protocols are particularly relevant for high-stakes domains such as medical diagnostics and legal reasoning, where wrong decisions can have real-world negative effects. We release the code and data for these experiments publicly[1].

Key Contributions:

§4.1 A systematic comparison of decision protocols, revealing for the first time that consensus is most effective for knowledge tasks, while voting is better in reasoning tasks.

§4.2 An analysis of scaling differences between increasing the number of agents and extending the number of communication rounds.

§4.3 Two new methods,AAD and CI, to enhance answer diversity in multi-agent debates by encouraging independent thinking.

[1] github.com/lkaesberg/decision-protocols

2 Related Work

LLMs as Agents. An agent differs from an LLM in that it has a defined planning behavior, can use tools, and maintains a state or memory across interactions. Techniques such as Chain-of-Thought (CoT) prompting (Wei et al., 2022), self-refinement (Madaan et al., 2023), and self-consistency (Wang et al., 2023) improve models' ability to plan, critique, and refine responses. Persona-based prompting (Jiang et al., 2024) enables LLMs to adopt specialized roles, improving answer diversity. A single-agent system operates as one entity with an internal state, while a multi-agent debate consists of multiple agents with a private state that persists across calls and may operate asynchronously, leading to emergent, independent behaviors (Du et al., 2023a; Zhao et al., 2023; Xu et al., 2023; Suzgun and Kalai, 2024; Goldberg, 2024). In multi-agent debates, many parameter choices have to be made, such as in which turn order they communicate (Yin et al., 2023), and which tools they can use (Yao et al., 2023). Yet one choice is inevitable: How to make decisions between agents.

Decision-Making. Finding collective solutions markedly impacts human decision-making (Jones, 1994). While consensus promotes shared decisions and allows everyone to contribute to the final solution, it can be time-consuming and lead to power concentration of individuals with "vetos". Conversely, voting can lead to a faster final decision because it streamlines decision-making, but is susceptible to manipulation and does not take into account all opinions. For example, participants can vote for a less preferred but more viable alternative to block an undesired outcome, leading to outcomes that do not reflect the collective will of the group (List, 2022).

Research in multi-agent debate has implemented various human decision protocols (Yin et al., 2023; Chen et al., 2023; Yang et al., 2024b). Exchange-of-Thought (Yin et al., 2023) employs a consensus-based approach, where agents iteratively refine answers through discussion in reasoning tasks, but they do not compare consensus to other decision protocols. Yang et al. (2024b,a) introduce multiple voting protocols (e.g., approval voting) but do not compare the performance of these voting decision protocols across different tasks, focussing more on a comparison of how humans vote compared to LLMs. ReConcile (Chen et al., 2023) integrates a hybrid voting and consensus approach

by iteratively refining answers through confidence-weighting until consensus is reached. In summary, prior work focuses on a single class of decision protocols without systematic comparison, partly because many parameters change between experiments (Yin et al., 2023; Yang et al., 2024b; Chen et al., 2023). This work systematically evaluates seven voting and consensus approaches on both knowledge and reasoning datasets to show task-based advantages of one protocol over another.

3 Methodology

In the following, we explain the multi-agent environment, decision protocols, response generators, and datasets used in our experiments.

3.1 Setup

We run multi-agent debates based on three key components: a discussion paradigm, a decision protocol, and an agent response generator. Each discussion consists of three automatically generated expert personas for the task following Kim et al. (2024). While the number of personas can vary, Yin et al. (2023) found this number to be the most efficient. Using personas is important as it generates agents with expertise in different domains, incorporating diverse viewpoints for the final decision. After that, the agents discuss the problem for multiple turns. The number of turns varies from one to five, depending on the experiment setup and decision protocol. The *decision protocol* defines what criteria must be met for the discussion to end and how the final solution will be created. Each agent can generate one answer per turn and, based on the agent's turn order, respond to the other agents as a default behavior. If not explicitly stated, all agents exchange messages with one another. These discussion characteristics are defined by the *discussion paradigm*. For consensus decision protocols, each agent also indicates whether they agree with the previous message. We repeat this process each round until a certain level of agreement (defined by the decision protocol) is reached, which leads to the final solution. For voting decision protocols, the agents start voting after the third turn. If they successfully agree on a solution in any turn after the third, the discussion ends. Agents can only memorize messages from up to two turns to limit the context length provided to the agent. We use a *response generator* to define how agents respond to previous messages. It determines how the discussion history is presented, what additional information is provided, how the persona is introduced, and in which tone they should respond (e.g., neutral or critical). Additional details on the experimental setup can be found in Appendix C.

The LLMs used for the experiments are Llama 3 8B and 70B[2] (META, 2024). Within one discussion, all prompted agents use the same base model. The smaller model uses two NVIDIA RTX5000 with 16 GB of VRAM (for $\sim$ 250 hours), and the larger model uses eight NVIDIA A100 with 40GB VRAM (for $\sim$ 40 hours). A list of experiment-specific parameters is available in Appendix F.

3.2 Agent Prompts and Decision Protocols

Prompt design has a marked impact on multi-agent debate. For this work, we propose three response generators. The **Simple Response Generator** is our default in which agents are prompted to answer neutral and unbiased to previous messages. The **Critical Response Generator** encourages agents to critically assess prior answers and propose new solutions, countering sycophantic tendencies (Sharma et al., 2023). The **Reasoning Response Generator** restricts agents to sharing only reasoning paths, thus avoiding bias towards agreeing with other agents' final solutions.

Decision protocols then determine the final solution of the discussion by selecting the most promising solution based on predefined mechanisms. We use two classes of protocols in this work.

Consensus decision protocols decide on an answer by prompting the agents to converge on one shared solution. The solution is selected when a required level of agreement among agents is reached. We explore three major agreement levels: *majority consensus* - more than 50% agreement, *supermajority consensus* - more than 66% agreement, and *unanimity consensus* - all agents have to agree. We extend the majority consensus of Yin et al. (2023) by higher agreement levels, i.e., supermajority (66%), unanimity (100%).

Voting decision protocols allow several possible solutions to be presented in parallel during the discussion, and ultimately all agents must vote on a final solution. If there is a tie in the voting, all agents will discuss for another round and then vote again. Our work includes four different voting decision protocols inspired by Yang et al. (2024b):

[2]The two models used for this study can be found at meta-llama/Meta-Llama-3-8B-Instruct and meta-llama/Meta-Llama-3-70B-Instruct

		Knowledge-Based			Reasoning-Based		
		MMLU (Accuracy)	MMLU-Pro (Accuracy)	GPQA (Accuracy)	SQuAD 2.0 (F1-Score)	StrategyQA (Accuracy)	MuSR (Accuracy)
	Baseline	44.8 ± 1.9	28.5 ± 1.3	28.9 ± 1.2	52.3 ± 2.2	51.7 ± 2.5	25.8 ± 1.6
	Baseline with CoT	53.1 ± 4.6	32.2 ± 4.7	29.9 ± 0.6	53.8 ± 1.2	55.5 ± 1.3	29.5 ± 2.6
Voting	Simple	53.3 ± 1.8	32.0 ± 2.7	30.5 ± 0.9	56.2 ± 0.5	58.5 ± 0.9	55.2 ± 1.5
Voting	Ranked	49.2 ± 1.5	33.1 ± 4.6	27.3 ± 3.9	58.0 ± 0.8	56.2 ± 3.4	52.5 ± 0.0
Voting	Cumulative	52.6 ± 4.0	28.3 ± 3.1	31.3 ± 2.8	55.8 ± 3.4	61.2 ± 1.6	56.8 ± 4.2
Voting	Approval	43.0 ± 2.1	29.2 ± 5.2	31.3 ± 2.6	46.5 ± 1.4	58.7 ± 0.4	50.9 ± 4.0
Voting	Average[3]	51.7 ± 2.4	31.1 ± 3.5	29.7 ± 2.5	56.7 ± 1.6	58.6 ± 2.0	54.8 ± 1.9
Consensus	Majority	53.2 ± 2.5	36.4 ± 2.1	32.3 ± 2.9	43.1 ± 2.1	59.9 ± 0.1	27.8 ± 2.5
Consensus	Supermajority	54.6 ± 3.6	35.2 ± 3.0	30.7 ± 2.1	44.4 ± 0.4	56.4 ± 2.1	29.3 ± 2.6
Consensus	Unanimity	54.2 ± 1.0	36.3 ± 0.4	30.0 ± 2.3	43.4 ± 2.0	58.8 ± 2.6	28.2 ± 2.8
Consensus	Average	54.0 ± 2.7	36.0 ± 1.8	31.0 ± 2.4	43.6 ± 1.5	58.4 ± 1.6	28.4 ± 2.6

Figure 2: Task performance$_{\pm \text{std}}$ for seven decision protocols (voting and consensus-based) on six tasks (knowledge and reasoning) based on agents with Llama 8B. **Bold** indicates the highest results per dataset. Standard deviation for three runs.

simple voting - each agent casts one vote, and the answer with the most votes wins; *ranked voting* - each agent ranks the solutions from best to worst to find the solution that consistently has the highest rank over all individual rankings; *approval voting* - each agent casts unlimited votes, and the answer with the most votes wins; and *cumulative voting* - each agent is allowed to divide up to 25 points between all solutions. The solution with the most points wins.

3.3 Datasets

We evaluate our decision protocols using six datasets: three knowledge tasks (i.e., MMLU, a broad-topic test; MMLU-Pro, a domain-specific test with challenging questions; GPQA, a specialized question set difficult for web search) and three reasoning tasks (i.e., StrategyQA, a multistep reasoning task; MuSR, a long-context murder mystery; SQuAD 2.0, a reading comprehension task with questions that may or may not have answers in the context). Because of computational constraints, we calculate the task performance on a subset of samples with three runs and report their standard deviation. More details about the datasets, the number of samples, and the sampling strategy can be found in Appendix B.

4 Experiments

We use a set of diverse reasoning and knowledge tasks to explore the effectiveness and limits of decision protocols in multi-agent debate.

[3]Approval Voting is left out as it consistently fails to reach a voting decision as described in Section 4.1.

4.1 Performance of Decision Protocols

Multi-agent debates for task solving have shown promise in recent research, but they rely on a varying choice of decision protocols, hindering systematic comparison (Yin et al., 2023; Chen et al., 2023). This experiment systematically compares a set of seven decision protocols, changing only the decision protocol used and nothing else. Specifically, we determine whether some protocols have advantages over others and how they behave in specific edge cases, such as for unanswerable questions.

We experiment with four voting-based (Simple, Ranked, Cumulative, Approval) and three consensus-based (Majority, Supermajority, Unanimity) methods and test them on three knowledge tasks (MMLU, MMLU-Pro, GPQA) and three reasoning tasks (SQuAD 2.0, StrategyQA, MuSR). We use the Llama 3 8B model with and without CoT prompting to generate baseline results for comparison with multi-agent debate. We compare answerable and unanswerable questions of the SQuAD 2.0 dataset to inspect how decision protocols behave in edge cases.

Figure 2 shows the baselines compared to multi-agent debate, with the decision protocols being the rows grouped by voting and consensus-based protocols. The columns are the datasets grouped by knowledge and reasoning-based tasks. Overall, multi-agent debate with consensus and voting decision protocols outperforms the CoT baseline. Notably, consensus protocols perform better on knowledge-based tasks with average improvements of 2.3% on MMLU, 4.9% on MMLU-Pro, and 1.3% on GPQA over the voting average. Voting

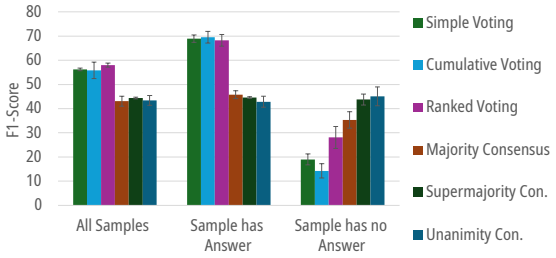

Figure 3: $F1_{\pm std}$ of different decision protocols on SQuAD 2.0 divided into three ablation groups: (middle) samples with an answer in the context (**Sample has Answer**), (right) samples with no answer in the context (**Sample has no Answer**), and (left) the combination of both (**All Samples**). Standard deviation over three runs.

protocols perform better in reasoning tasks with average improvements of 13.1% on SQuAD 2.0, 0.2% on StrategyQA, and 26.4% on MuSR over the consensus average. The results for the Llama 3 70B model can be found in Appendix A.1. On average, voting-based protocols require 3.38 rounds and consensus-based protocols 1.42 rounds to reach a decision, indicating that consensus is more suitable for applications that require quick decision-making as they use less test-time compute. Interestingly, most conversations reach a decision before they are terminated after five turns. Approval voting is the only protocol that often leads to no agreed solution (59% of cases) due to the agreeableness of the agents. The agents often simply vote for all answers as they do not have to decide, leading to many ties (see Appendix A.2 for details). Figure 3 shows the F1-Score for SQuAD 2.0 divided into whether questions are answerable or not for each decision protocol. Voting achieves a higher overall F1-score (56.7%) than consensus (43.6%), especially for answerable samples. However, consensus is more effective for unanswerable samples, and stricter consensus methods (e.g., unanimity consensus) produce even higher scores.

The effectiveness of decision protocols depends on the nature of the task. Our results suggest that consensus excels in knowledge tasks, and voting works better for reasoning tasks. Voting protocols might outperform consensus in reasoning tasks because they allow the exploration of multiple reasoning paths. One reason for the good performance of consensus in knowledge tasks is that it mitigates individual agent errors by requiring multiple agents to agree on the same statement before reaching a conclusion. This approach is less susceptible to manipulation or errors by any single agent. We found voting can be easily tricked by generating answers that seem to be correct but have some incorrect statements (see Appendix G.2). The consensus method has a distinct advantage by incorporating a safeguard of repeated checks across agents to find these small errors. The detailed analysis of SQuAD 2.0 further supports the finding that decision protocol effectiveness is task-dependent.

Both experiments reveal that the decision protocol used can significantly impact the overall performance of multi-agent debates. While multi-agent approaches generally require more computation than a single agent (approximately five times the compute resources for consensus protocols and ten times for voting protocols compared to the CoT baseline in our setup), the choice between protocols offers distinct advantages beyond simply scaling resources. Notably, consensus protocols often reached a decision faster than voting protocols, particularly on knowledge tasks that also achieved superior performance (Figure 2). Conversely, voting excelled on reasoning tasks despite potentially needing more rounds and compute. This demonstrates that with the correct choice, multi-agent debates can be made more error-resistant by using consensus decision protocols and more explorative by using voting decision protocols. However, we acknowledge that these performance improvements must be weighed against the substantial increase in computational resources required compared to simpler baseline methods. High token consumption is a significant practical limitation of the multi-agent debate (MAD) paradigm, and its prevalence in the field does not diminish its importance as a challenge. Observed performance improvements, although positive, may not always scale proportionally with the increased computational resources. Our study's primary objective was to systematically compare how different decision protocols function within this resource-intensive MAD paradigm to help researchers select the most effective protocol for their specific tasks. Therefore, researchers should carefully consider whether the incremental performance gains justify the associated computational costs in their particular use cases. An additional experiment comparing solution counting (selecting the answer that is given the most) to our prompted decision protocols showed similar trends and had little impact on our main findings (see Appendix D).

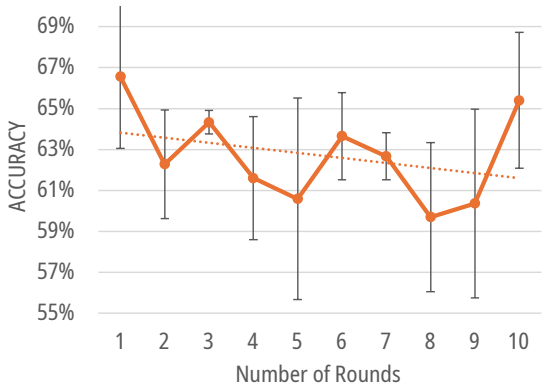

Figure 4: Accuracy$_{\pm std}$ on StrategyQA when the agents have to talk for a given number of rounds before they are allowed to vote using the simple voting decision protocol. Standard deviation over three runs.

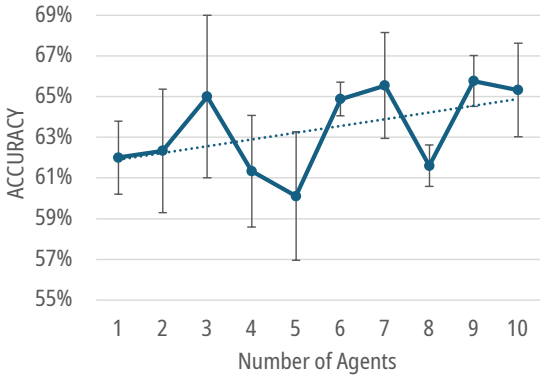

Figure 5: Accuracy$_{\pm std}$ on StrategyQA with a different number of agents participating in the discussion. The final answer is created using the simple voting decision protocol. Standard deviation over three runs.

4.2 Number of Agents and Discussion Rounds

Research by Yin et al. (2023) suggests that increasing the number of rounds and agents participating in a multi-agent debate is beneficial because of test-time compute scaling. In our experimental setup, agents communicate for three rounds before voting, following Du et al. (2023b). As our first experiment showed, consensus reaches a decision much faster than voting. Here, we investigate how the number of discussion rounds before voting impacts task performance, as fewer rounds with the same accuracy would improve computational efficiency. Becker (2024) suggests that extended discussions may lead to decreased accuracy because of agents drifting away from the original task. Wang et al. (2024a) shows increasing the number of agents may have positive effects.

We conduct this experiment using the StrategyQA dataset because multi-agent debate has consistently mitigated errors of single agents using CoT as shown in the previous experiment. The experiment is structured in two parts. First, we fix the number of agents to three and increase the number of discussion rounds from one to ten. Second, we fix the number of rounds to three and increase the number of agents from one to ten. Both experiments use the *simple voting* decision protocol, and each condition is evaluated across three independent runs.

Figure 4 shows accuracy relative to the number of rounds the agents have to communicate before voting starts. A linear regression, visualized by the dotted line, indicates that increasing discussion rounds slightly reduces accuracy. Still, with the increase from nine to ten rounds a rather abrupt increase in task performance can be seen. Separately, Figure 5 shows the task performance relative to the number of agents participating in the debate. Here, a linear regression indicates that increasing the number of agents leads to a slight upward trend in accuracy, suggesting that larger groups improve task performance.

Having more agents generating solutions generally leads to higher accuracy. This resembles the effect seen in self-consistency (Wang et al., 2023), where increasing the number of sampled answers improves task performance. However, Figure 4 shows that increasing the number of debate rounds decreases performance, contradicting the expected benefits of self-refinement. This aligns with recent concerns about the reliability of self-refinement, as Huang et al. (2023) questions the findings of Madaan et al. (2023), suggesting that the claimed improvements may not be robust.

At the same time, we also have one case where we see an increase in task performance as the number of rounds increases. Observe the sharp upside arc in Figure 4 for turn ten. To test whether individual rounds can recover this performance decrease, we create an ablation experiment where agents challenge the final result and propose a new solution. This extra round allows us to verify the effects of an extended discussion and also if agents would change their final solution. We create five different challenge scenarios and test them on two knowledge tasks (MMLU and MMLU-Pro) and two reasoning tasks (StrategyQA and SQuAD 2.0). In the first setting, we provide agents with the final solu-

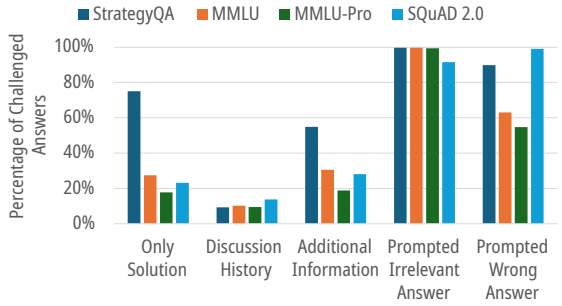

Figure 6: Percentage of agents challenging the final solution with different levels of information.

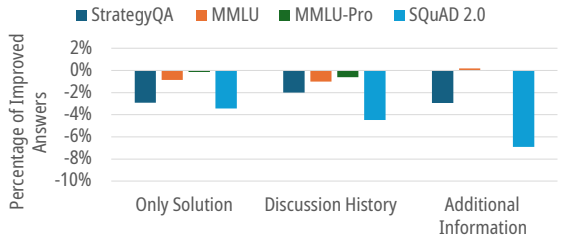

Figure 7: Percentage of answers that improve after being challenged using different levels of information.

tion and ask whether they agree or want to change it. In the second setting, agents can access the full discussion history with prior reasoning. The third setting includes additional context we retrieve with RAG from Wikipedia[4]. The fourth and fifth settings act as a baseline, introducing intentionally incorrect solutions—either irrelevant or wrong answers generated by the Qwen2 7B (Research, 2024) model to test whether agents can detect incorrect solutions. Solutions are generated with a different model than Llama 3, to remove any bias that might occur if the same model used to detect the wrong answer also generated it (Wahle et al., 2021).

We find that agents are less likely to challenge the final result when the discussion history is provided (10% decrease), as seen in Figure 6. For the StrategyQA dataset, this change is more prominent, with a 60% decrease in the challenge rate. Providing additional information has no noticeable effect compared to just providing the solution. Agents can detect irrelevant solutions (99% challenge rate) and incorrect solutions (75% challenge rate) with high accuracy. Figure 7 shows the percentage of challenged answers with a changed final score based on the new solution. It can be seen that this has no positive effect, especially on the reasoning tasks (StrategyQA and SQuAD 2.0), where it leads to worse solutions in 3% of cases. There are no major effects on knowledge tasks (MMLU and MMLU-Pro). Overall, the different challenge scenarios (Only Solution, Discussion History, and Additional Information) do not significantly impact the quality of the challenged answers, as they all perform very similarly.

These findings further support that more debate rounds harm task performance, suggesting the expected benefits of self-refinement may be unreliable and also differ from task to task. Therefore, we recommend scaling the number of agents as opposed to turns to create a larger knowledge base.

4.3 Answer Diversity

In multi-agent debates, answer diversity plays an important role in improving decision-making and task performance. A more diverse set of answers is beneficial because selecting the correct solution from multiple options is often easier than relying on a single agent's solution (Zheng et al., 2023). This principle is also key to self-consistency approaches (Wang et al., 2023) and explains why multiple-choice tests are often easier than open-ended ones. The goal of this experiment is to exploit this property and explore ways to increase answer diversity to optimize task performance.

In typical multi-agent debates, the first agent starts generating a possible solution for the given task. The next agent can either propose a new solution or improve on the previous solution. Throughout our experiments, we observe that agents are agreeable and often only improve the answer from the first agent without proposing an idea based on their own expertise. While the idea of independent answer generation has been previously explored in different multi-agent settings (Du et al., 2023b; Liang et al., 2024), these studies did not systematically formalize and evaluate it separately from other multi-agent discussion parameters. To address these limitations, we explicitly formalize this principle as All-Agents Drafting (AAD), a method that forces each agent to generate an independent solution based on their own expertise in the first round. Second, we propose Collective Improvement (CI) which starts similarly to AAD with each agent generating an independent solution, but unlike AAD where each agent can communicate with another after the first turn, CI prohibits communication and only shows the solutions from the previous turn to the agent in the next turn. The rationale behind this is that agents often immediately get biased by other proposals rather than fol-

[4]We use: github.com/Multi-Agent-LLMs/context-plus

Strategy	Baseline	All draft	Collective	Critical	Reasoning
Answer Cosine Similarity	0.888	0.870	0.845	0.843	0.916
Mean Accuracy	58.3%	62.8%	65.7%	59.4%	51.9%
Delta Baseline	0.0%	3.3%	7.4%	1.1%	-6.4%

Table 1: Final answer similarity based on average cosine similarity between SBERT embeddings of final answers compared to task performance on StrategyQA dataset.

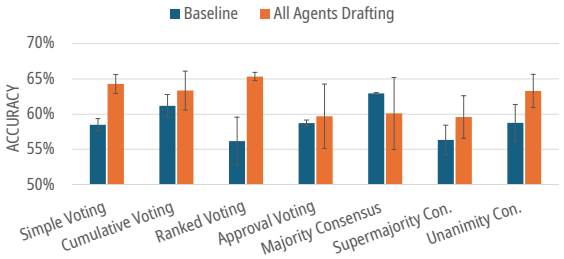

Figure 8: Comparison of agents iterating on one initial draft (baseline) vs. each of the three agents generating an initial draft separately on the StrategyQA dataset using AAD. Standard deviation over three runs.

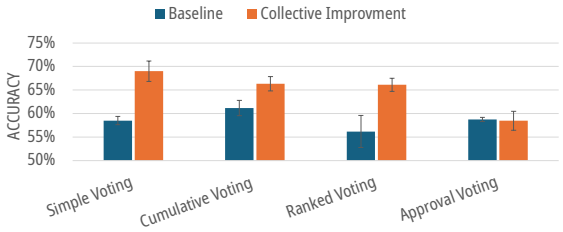

Figure 9: Comparison of agents discussing using the default discussion protocol compared to the CI discussion protocol on the StrategyQA dataset. Standard deviation over three runs.

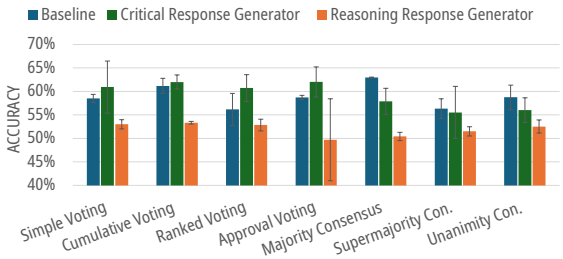

Figure 10: Comparison of agents using the default response generator compared to the reasoning and critical response generator on the StrategyQA dataset. Standard deviation over three runs.

lowing their own ideas which leads to less answer diversity. Thus, our new proposal, CI, extends the independent generation method with a discussion paradigm designed explicitly for voting protocols, demonstrating improved exploration and diversity of ideas in the multi-agent setting. CI only works for voting decision protocols as it removes the turn order of the agents, and therefore, no iterative improvements and agreements can be made. Third, we introduce a *critical response generator* and a *reasoning response generator*. Both test whether it is possible to improve answer diversity by prompting agents to respond more critically or by allowing them to exchange only reasoning steps and no final solutions during the discussion to avoid agreeableness towards final solutions and not reasoning ideas. To quantify if these methods increase answer diversity, we calculate the cosine similarity between the SBERT embeddings (Reimers and Gurevych, 2019) of the agent's answers and correlate it with changes in task performance.

Figure 8 shows the accuracy of AAD compared to the multi-agent baseline on StrategyQA for all decision protocols. AAD increases the performance on average by 3.3% over the baseline. The strongest improvements are for simple and ranked voting. Cumulative voting, supermajority, and unanimity consensus perform a bit worse. Approval voting and majority have no real improvement but are within the standard deviation of the baseline.

Figure 9 shows the accuracy of CI compared to the multi-agent baseline. CI is a voting–only extension. On average, CI performs 7.4% better than the baseline.

Figure 10 shows the accuracy of different response generators compared to the baseline for different decision protocols. The results show that the *critical response generator* does not provide a reliable improvement, while the *reasoning response generator* even leads to a decrease in task performance.

Table 1 shows the cosine similarity, mean accuracy, and delta in performance as the rows for the different improvement methods (AAD, CI and *response generators*) as columns. We observe a general trend that if the answer diversity increases, task performance increases, too. The *reasoning response generator* decreases answer diversity, which results in a drop in task performance.

AAD and CI achieve higher answer diversity

and, therefore, higher task performance, but the two response generators struggle to provide reliable results and may even reduce task performance. The critical and restrictive prompting style degrades the quality of the discussion by forcing a specific answer style that is not always beneficial. Therefore, we recommend using our methods AAD or CI to limit group interactions and promote independent thinking but caution against using methods that directly change the response behavior of the agents, as this may have unwanted side effects.

5 Conclusion

We systematically evaluated the role of consensus and voting decision protocols across three knowledge and three reasoning tasks. Our study assessed how the number of discussion rounds and agents influences task performance. We proposed two new methods to improve answer diversity during multi-agent discussions and decisions, i.e., All-Agents Drafting (AAD) and Collective Improvement (CI). AAD requires each agent to contribute draft ideas at the beginning of the discussion, and CI encourages independent reasoning steps by limiting communication between agents and only allowing them to exchange possible solutions after each turn.

Our findings show that voting performs well on reasoning tasks, outperforming consensus by up to 13.2%, and outperforming a single CoT baseline by 10.4%. This is likely because voting-based protocols allow agents to explore multiple reasoning paths instead of a single one, as in consensus. In comparison, consensus outperforms voting in knowledge tasks by up to 2.8%, because it improves fact-checking by requiring at least the agreement of the majority of agents. Increasing the number of agents in the discussion improved task performance, while increasing the number of discussion rounds before voting decreased performance. One possible reason for that could be problem drift, a recent limitation found in long multi-agent debate (Becker, 2024). AAD improved performance by up to 3.3%, and CI by up to 7.4% over multi-agent debate baselines, and 6.1% and 10.2% over single model CoT baseline respectively. Our methods enhance answer diversity and reveal a connection between answer diversity and task performance.

Future work could explore other characteristics influencing decisions, such as power relations between managers and employees. This could also involve examining personas within this hierarchical structure to investigate whether dominant or affectionate leaders are more effective in leading discussions (Ames, 2009).

We recommend using voting in reasoning tasks and consensus in knowledge tasks, scaling up the number of agents instead of the number of rounds, and increasing the diversity of answers between agents using AAD and CI.

Limitations

Multi-agent debates are computationally expensive because they require a message from each agent in each round, quickly leading to hundreds of forward passes per model. Because of the high computational cost and the range of decision protocols and tasks in our work, we used sampled subsets of the datasets, which can lead to some variance. To control for that variance, we sampled with a 95% confidence level and calculated the standard deviation of three independent runs.Overall, the results were markedly higher than what could be explained by the standard deviation. More details about the dataset and other parameters can be found in Appendix B. Despite efforts to improve answer diversity, agents often converged on similar responses, suggesting that more advanced techniques to encourage independent solutions are needed in the future.

While this study focuses on decision protocols, we acknowledge that prompt design and persona selection are also relevant parameters. Initial explorations with more diverse prompt structures did not yield consistent improvements over the simpler, reproducible design presented. Furthermore, Section 4.3 analyzes variations in agent response styles (critical, reasoning only), showing decision protocol dependent effects. A dedicated persona ablation was omitted due to significant computational cost and prior evidence suggesting limited impact on task performance in similar setups (Zheng et al., 2024).

Acknowledgments

This work was partially supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation. Many thanks to Andreas Stephan, Florian Wunderlich, and Niklas Bauer for their thoughtful discussions and feedback.

References

Daniel Ames. 2009. Pushing up to a point: Assertiveness and effectiveness in leadership and interpersonal dynamics. *Research in Organizational Behavior*, 29:111–133.

Jonas Becker. 2024. Multi-Agent Large Language Models for Conversational Task-Solving.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs.

William G. Cochran. 1953. *Sampling techniques*. Sampling techniques.

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023b. Improving Factuality and Reasoning in Language Models through Multiagent Debate.

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*.

Yoav Goldberg. 2024. Are multi-llm-agent systems a thing? yes they are. but.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large Language Models Cannot Self-Correct Reasoning Yet.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Bernie Jones. 1994. A comparison of consensus and voting in public decision making. *Negotiation Journal*, (2).

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a Double-edged Sword: Mitigating the Negative Impact of Role-playing Prompts in Zero-shot Reasoning Tasks.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models.

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.

Christian List. 2022. Social Choice Theory. In *The Stanford Encyclopedia of Philosophy*, winter 2022 edition.

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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

META. 2024. The Llama 3 Herd of Models.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark.

Alibaba Research. 2024. Qwen2 Technical Report.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2023. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding. *arXiv preprint*. ArXiv:2401.12954 [cs].

Steven K. Thompson. 2012. *Sampling*, 3rd ed edition. Wiley series in probability and statistics.

Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. 2021. Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, Norman Meuschke, and Bela Gipp. 2023. Ai usage cards: Responsibly reporting ai-generated content. *Preprint*, arXiv:2303.03886.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024a. 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)*.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark (Published at NeurIPS 2024 Track Datasets and Benchmarks).

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024c. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. ExpertPrompting: Instructing Large Language Models to be Distinguished Experts. *arXiv preprint*. ArXiv: 2305.14688 [cs].

Joshua C. Yang, Carina I. Hausladen, Dominik Peters, Evangelos Pournaras, Regula Hnggli Fricker, and Dirk Helbing. 2024a. Designing digital voting systems for citizens: Achieving fairness and legitimacy in participatory budgeting. *Digit. Gov.: Res. Pract.*, 5(3).

Joshua C. Yang, Marcin Korecki, Damian Dailisan, Carina I. Hausladen, and Dirk Helbing. 2024b. LLM Voting: Human Choices and AI Collective Decision Making.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

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*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2023. ExpeL: LLM Agents Are Experiential Learners. *arXiv preprint*. ArXiv:2308.10144 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Additional Results

Additional results for the first experiment to provide further information.

A.1 Task Performance with Llama 3 70B

Compared to the results of the Llama 3 8B model, the larger Llama 3 70B model performs much better overall, as seen in Figure 11. Most of the results are a bit better than the baseline, but the multi-agent discussions are only in a few cases able to outperform the CoT baseline. This model does not fail to follow the prompt for the MuSR baseline and consensus-based decision protocols. Therefore, the big performance gain from the smaller model cannot be observed here. SQuAD 2.0 and StrategyQA had the largest performance gains, even outperforming the CoT baseline, similar to the results from the smaller model. This difference in task performance can have many reasons. As Li et al. (2024) showed, smaller models are more likely to hallucinate, which reduces task performance. This can be mitigated by using multiple agents because it is less likely that two agents hallucinate the same things. Larger models tend to hallucinate less, reducing this effect for the Llama 3 70B model (Li et al., 2024). In general, Llama 3 70B has a much higher baseline for task performance, making it more difficult to improve baseline results. Many of the improvements by the Llama 3 8B model are quite small, except for the ones where the Llama 3 70B model also outperforms the CoT baseline. This can be taken as evidence that these multi-agent discussions require specific problem structures, or else the agents are just talking about the same results for multiple rounds and agreeing with each other. If these discussions continue too long, they can drift away from the original task, which reduces task performance. This has also been observed by Becker (2024) and an example can be seen in Appendix G.3. A positive example of how discussion can help task performance can be seen in Appendix G.1.

A.2 Termination percent

The data in Table 2 shows the number of turns that are needed for each decision protocol to reach a final decision for the MMLU dataset. Most of the voting decision protocols are able to vote for a final answer already in the first round in which they are allowed to vote. Simple voting has the highest agreement rate, but also cumulative and ranked voting only need in a few cases another round. In contrast, the approval decision protocol only achieves this in $\sim 27\%$ of the cases. About 14% need another round and the rest is canceled after the fifth round. This happens because these models like to agree with each other, and therefore they tend to vote for many of the answers, which often leads to a tie. Therefore, more restrictive voting decision protocols can reach a decision more easily, as a tie is less likely. The consensus decision protocols require only one to two rounds to reach consensus and still achieve a higher task performance because these results are based on the MMLU dataset. The decision protocols tend to behave similarly in terms of rounds needed to create a final answer, independent of task and model.

B Additional Details on Datasets

The dataset selection is very important for this work. It needs to be tested whether decision protocols perform well in multiple domains and whether some protocols perform better with specific tasks than others. Therefore, we selected datasets from different domains and divided them into two groups:

- **Knowledge-based Datasets**: MMLU, MMLU-Pro, and GPQA. These still require some reasoning and domain knowledge.
- **Reasoning-based Datasets**: StrategyQA, MuSR, and SQuAD 2.0. These emphasize multistep reasoning and textual comprehension.

An overview of all these datasets can be found in Table 3 with a description and the number of samples used for evaluation.

B.1 Sampling Strategy

Because multi-agent discussions are expensive, we use a small subset of each dataset that still represents the dataset effectively. This follows approaches used by Yin et al. (2023); Chen et al. (2023); Becker (2024) and ensures a 95% confidence level with a 5% margin of error:

$$n_0 = \frac{Z^2 \cdot p \cdot (1-p)}{d^2},$$

where $Z = 1.96$, $p = 0.5$, and $d = 0.05$ (Thompson, 2012). For finite datasets, a finite pop-

		Knowledge-Based			Reasoning-Based		
		MMLU (Accuracy)	MMLU-Pro (Accuracy)	GPQA (Accuracy)	SQuAD 2.0 (F1-Score)	StrategyQA (Accuracy)	MuSR (Accuracy)
	Baseline	72.7 ±0.8	57.5 ±0.7	45.2 ±2.0	69.5 ±0.8	76.6 ±2.3	63.2 ±1.2
	Baseline with CoT	75.5 ±0.8	60.8 ±1.0	45.9 ±1.5	68.0 ±1.4	78.2 ±1.0	66.8 ±1.6
Voting	Simple	75.5 ±1.3	56.5 ±5.4	45.7 ±0.3	69.5 ±0.5	81.2 ±1.4	59.3 ±3.0
Voting	Ranked	73.3 ±1.9	54.3 ±0.9	44.2 ±2.1	70.6 ±1.1	80.3 ±1.0	59.5 ±0.5
Voting	Cumulative	72.5 ±1.9	53.0 ±0.4	43.9 ±3.1	69.7 ±1.1	80.0 ±0.8	60.0 ±1.7
Voting	Approval	50.2 ±2.1	36.3 ±3.7	33.0 ±3.4	24.3 ±5.0	46.4 ±13.9	49.1 ±12.4
Voting	Average[6]	73.8 ±1.7	54.6 ±2.2	44.6 ±1.8	69.9 ±0.9	80.5 ±0.7	59.6 ±1.7
Consensus	Majority	74.0 ±1.4	57.3 ±3.0	43.7 ±1.0	58.2 ±1.0	80.1 ±0.3	61.3 ±3.3
Consensus	Supermajority	71.9 ±1.2	57.0 ±1.7	46.6 ±0.8	54.3 ±1.9	80.3 ±1.3	60.2 ±0.3
Consensus	Unanimity	72.2 ±2.4	57.3 ±1.5	45.3 ±2.5	56.7 ±2.2	78.1 ±2.3	61.3 ±0.8
Consensus	Average	72.7 ±1.7	57.2 ±2.1	45.2 ±1.4	56.4 ±1.7	79.5 ±1.3	60.9 ±1.5

Figure 11: Task performance$_{\pm \text{std}}$ for seven decision protocols (voting and consensus-based) on six tasks (knowledge and reasoning) based on agents with Llama 70B. **Bold** indicates the highest results per dataset. Standard deviation for three runs.

Group	Turn 1	Turn 2	Turn 3	Turn 4	Turn 5[5]	Task Performance Score
Voting	0.00%	0.00%	99.33%	0.50%	0.17%	53.3 ± 1.8
Cumulative	0.00%	0.00%	94.00%	5.50%	0.50%	52.6 ± 4.0
Ranked	0.00%	0.00%	91.17%	7.83%	1.00%	49.2 ± 1.5
Approval	0.00%	0.00%	26.67%	14.33%	59.00%	43.0 ± 2.1
Majority	80.00%	13.67%	4.83%	1.00%	0.50%	53.2 ± 2.5
Supermaj.	79.33%	14.33%	4.83%	1.00%	0.50%	54.6 ± 3.6
Unanimity	59.50%	21.67%	12.67%	3.50%	2.67%	54.2 ± 1.0

Table 2: Number of rounds needed for each decision protocol to reach a final decision for the MMLU dataset with Llama 3 8B.

ulation correction is applied:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}},$$

where N is the total number of samples in each dataset (Cochran, 1953). The specific sample sizes reflecting this calculation are listed in Table 3.

B.2 Repeatability

Each dataset was tested three times to obtain a standard deviation of the results (Reimers and Gurevych, 2017; Chen et al., 2023; Becker, 2024), ensuring reliable and robust performance estimates across multiple evaluations.

C Multi-Agent Framework

For our experiments, we use the **M**ulti-**A**gent **L**arge **L**anguage **M**odels (MALLM) framework[7].

[5]In this round, the discussion is terminated.

[6]Approval Voting is left out as it consistently fails to reach a voting decision as described in Section 4.1.

[7]Available here: github.com/Multi-Agent-LLMs/mallm

C.1 Architecture Overview

To better understand the different modules, we take a closer look at each component and what role it plays in creating multi-agent discussions. An overview can be found in Figure 12 as it provides an example workflow for the framework and how a discussion is created. The discussion starts with generating personas relevant to the given task and assigning them to the participating agents. The personas are generated using the same LLM which is later used for the agents. After that the agents start to generate solutions and improve the suggestions from the other agents. The turn order of the agents is defined by the *discussion paradigm*. This also defines which answers are visible to other agents and who can talk to whom. The *response generator* defines how an agent receives the other answers and also the way it responds. After a certain number of rounds or when enough agents agree, a *decision protocol* is used to select the best answer either via voting or just by looking for a certain consensus threshold. If the decision protocol fails, for exam-

Dataset	Description	Samples	Eval-Samples
Knowledge-based			
MMLU	Massive Multitask Language Understanding benchmark covering 57 subjects	14,042	375 (x3)
MMLU Pro	Professional-level extension of MMLU with advanced questions	12,032	374 (x3)
GPQA	Challenging dataset of multiple-choice questions written by domain experts in biology, physics, and chemistry	546	250 (x3)
Reasoning-based			
StrategyQA	Dataset of questions requiring implicit multi-hop reasoning	2,289	330 (x3)
MuSR	Logic reasoning for solving murder mystery stories	250	152 (x3)
SQuAD 2.0	Stanford QA Dataset with answerable and unanswerable questions	11,873	373 (x3)

Table 3: All datasets used for evaluation with a short description, number of samples in the test set, and number of samples used in this study. The datasets are divided into knowledge-based tasks and reasoning-based tasks.

ple, due to a tied vote, the discussion continues for another round. In the framework a parameter can be defined to terminate discussions after a certain number of rounds to make sure they do not communicate forever.

Agent Personalities. The first step of the discussion is the generation of agent personas. Each of the agents participating in the discussion has a certain persona assigned to them. This can unlock more knowledge for the LLM on a specific topic (Kim et al., 2024). To get the best results, we want as diverse personas as possible while still maintaining them to be relevant to the task. The default setting for the framework is to prompt a LLM and ask for a persona relevant to the given task (Wang et al., 2024c). After each generation it also provides the generated personas to avoid duplication. This way of generating personas provides a good starting point, but as this is built as a modular component, it can be swapped out with another function, which, for example, generates half of the agents with this method and initializes the other ones as neutral agents without a persona.

Response Generators. Another important part of multi-agent discussions is how each agent responds to the previous responses. Do we use CoT to improve performance, or does this result in too long answers? By changing the way an agent is prompted, a lot of performance can be gained or lost. Therefore, it is key to make this as customizable as possible. The researcher has the possibility to change the default behavior (neutral answers), for example, by prompting the agent to be more critical or changing the way the discussion history is presented. The system prompt for the agent's persona can also be adjusted. MALLM already has many different built-in response generators. The ones relevant for this work are the following.

- **Free Text** is the most basic form of the agent prompt. Each agent gets a predefined number of discussion history rounds as memory. The prompt language is neutral, and the task is presented each round to mitigate the potential drift from the topic of the discussion (Becker, 2024). In addition, the agent is always asked to agree or disagree with the answer of the previous agent.

- **Simple** behaves very similar to the Free Text response generator, but the prompt is a bit simpler to make it easier to understand for the LLM and reduce the context length.

- **Critical** forces the agent to respond very critically to the previous answer and try to find new solutions. Some studies have shown that LLMs can show some form of sycophancy, which is not helpful for a constructive discussion (Sharma et al., 2023). Encouraging them to be more critical may reduce this.

- **Reasoning** doesn't allow the agents to com-

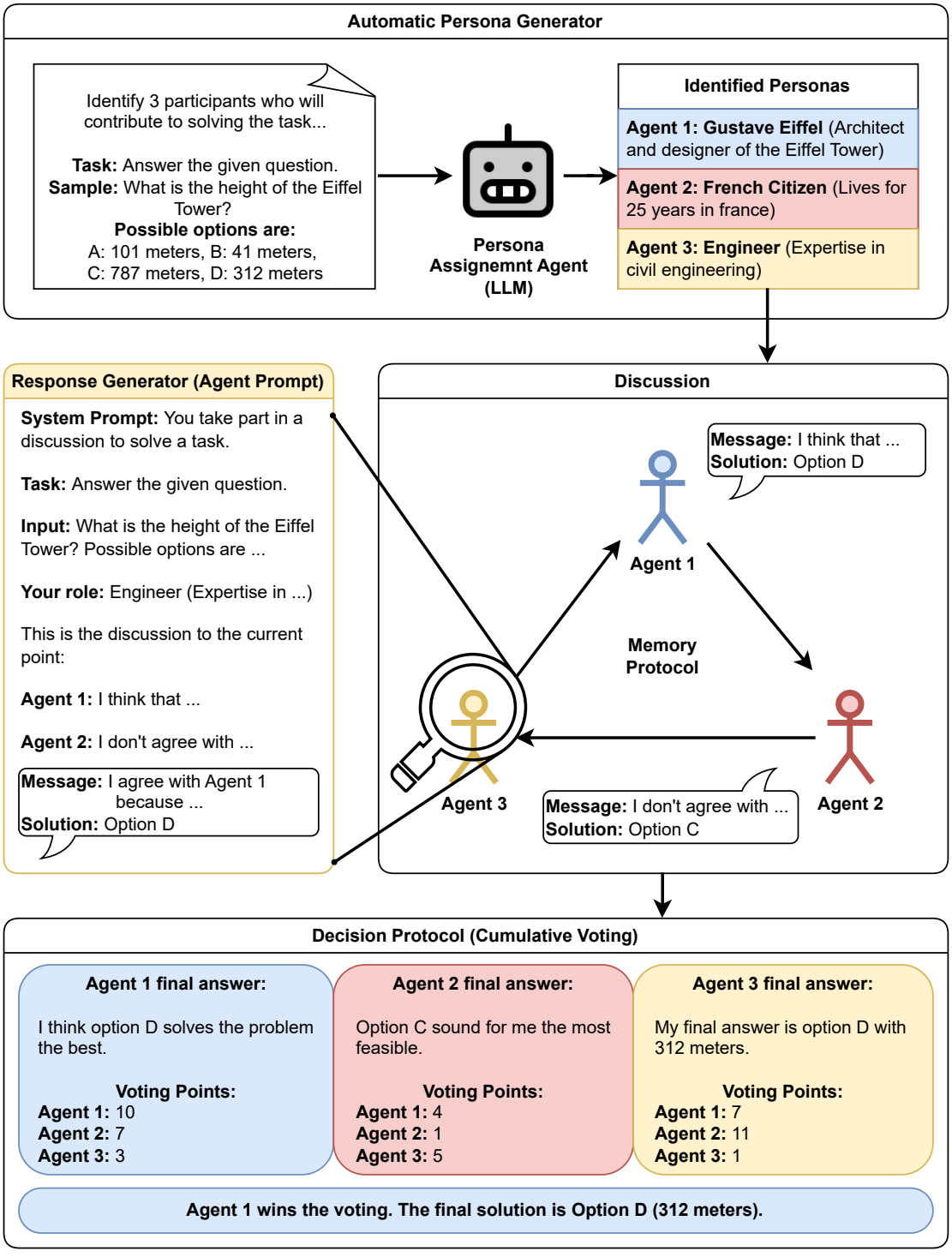

Figure 12: Example multi-agent discussion conducted in the MALLM framework. It showcases the functionality of the four modules and how they work together to get an improved final solution.

municate their final solution with the other agents. They can only share reasoning that can be used to find a final solution. In the end, each agent has to come up with its own solution without being directly influenced by other agents.

Discussion Paradigms. These paradigms define the discussion format for the entire task. They can control the order in which the agents communicate with each other, and which answers are visible only to certain agents. Currently, all the built-in discussion paradigms are static, meaning that the turn order is predefined and cannot be changed based on specific events during the discussion. However, due to the modular nature of MALLM, a new discussion paradigm can be added, for example using an LLM as a moderator to dynamically decide which agent should respond next. Current research by Yin et al. (2023) and Becker (2024) shows that discussion protocols have little impact on downstream task performance. MALLM includes the following discussion paradigms, which are illustrated in Figure 13. The first four paragdims are inspired by the work of Yin et al. (2023), while the fifth was developed as part of this work.

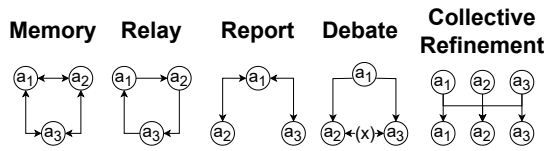

Figure 13: Illustration of Discussion Paradigms available for use in MALLM

- **Memory** is the most basic discussion paradigm. The agents respond to the solution of the previous agents with feedback or an improved solution. All answers are visible to the other agents.

- **Relay** behaves similarly to the memory paradigm. The turn order is the same, but each agent can only see the answer from the previous agent.

- **Report** introduces one agent as a moderator that can communicate with other agents. The other agents can only communicate with the moderator and have access to these messages only. Only the moderator can see all messages.

- **Debate** is similar to the report paradigm, as it also needs a moderator. Here, the other agents can communicate for a predefined number of rounds before they forward their reasoning to the moderator agent, which starts the next round of debate.

- **Collective Refinement** In this protocol, each agent first generates an answer independently. In each subsequent round, every agent receives the responses from all other agents at the same time. Using this shared information, each agent refines their own answer. This process continues throughout the rounds, helping agents gradually reach a shared and improved solution. There is no turn order, and all agents have the same level of knowledge in each round.

Decision Protocols. These are crucial for the framework as they decide which answer gets presented as the final answer to the problem. Multi-agent discussions produce multiple results for the same problem because each agent has its own reasoning and ideas on how to solve the problem. Therefore, some process is needed to decide which answer looks the most promising. We divide these decision protocols into three subtypes that we want to analyze. An overview of how each of these decision protocols works theoretically can be found in Figure 1 and all prompts used for them can be found in Appendix E.

Consensus Based Decision Protocols. These are the simplest kinds of decision protocols. After each answer, the next agent has to agree or disagree with the previous statement. Depending on the response generator, this happens in the same message, and the agreement is extracted with a regular expression, or this is split into multiple answers. If enough agents agree in order, there is a consensus. The final answer is extracted by instructing the last agent to solve the given task with the information available in the latest messages. The prompt used for this can be found in Appendix E.1. There are several types of consensus decision protocols available in MALLM. **Majority consensus** requires 50% of the agents to agree. **Supermajority consensus** requires 66% of the agents to agree, and **unanimity consensus** requires all agents to agree.

Voting Based Decision Protocols. For voting based decision protocols, the process differs

Decision Protocol	MMLU	MMLU-Pro	GPQA	StrategyQA
Baseline				
Solution Counting	53.6 ± 2.8	33.0 ± 3.5	28.6 ± 2.8	55.3 ± 1.6
Voting Protocols				
Simple Voting	53.3 ± 1.8	32.0 ± 2.7	30.5 ± 0.9	58.5 ± 0.9
Ranked Voting	49.2 ± 1.5	33.1 ± 4.6	27.3 ± 3.9	56.2 ± 3.4
Cumulative Voting	52.6 ± 4.0	28.3 ± 3.1	31.3 ± 2.8	**61.2 ± 1.6**
Consensus Protocols				
Majority Consensus	53.2 ± 2.5	**36.4 ± 2.1**	**32.3 ± 2.9**	59.9 ± 0.1
Supermajority Consensus	**54.6 ± 3.6**	35.2 ± 3.0	30.7 ± 2.1	56.4 ± 2.1
Unanimity Consensus	54.2 ± 1.0	36.3 ± 0.4	30.0 ± 2.3	58.8 ± 2.6

Table 4: Comparison of solution counting with prompted decision protocols across four datasets. Performance is Accuracy ± std. **Bold** indicates the highest result per dataset.

slightly compared to consensus-based decision protocols. The agents are forced to discuss for a predefined number of turns and afterward create a final solution. In the default setting, they have to discuss for three rounds, as current research such as Du et al. (2023b) shows that this allows for reasonable strong improvements considering computing resources. If there happens to be a tie in the voting, the agents have to discuss it for another round, and after that, they are asked to vote again. If they do not reach a final decision before exceeding the maximum number of rounds (defined in the discussion configuration), the solution of the first agent is used. To analyze the impact of the voting procedure, different processes similar to the work of Yang et al. (2024b) are implemented.

- **Simple Voting** Each of the agents has only one vote. They can vote for any other agent or for themselves. The agent with the most votes wins.

- **Ranked Voting** The agents have to rank all final answers. The best solution is chosen by adding the ranking indices for a given agent and then selecting the answer with the best cumulative rank.

- **Cumulative Voting** Each agent has to distribute up to 25 points to all possible answers. They can also give fewer points and freely divide the points between all agents (even themselves). The winner is selected by adding all the points for a given agent and selecting the final answer with the most points.

- **Approval Voting** The agent has to provide a list of solutions that it approves. After that, the approvals from all agents are counted, and the answer with the most approvals wins the vote.

D Solution Counting Ablation

We conducted an ablation study comparing solution counting with our prompted decision protocols to investigate the effectiveness of simpler voting methods. Solution counting chooses the answer the agents give most frequently as the final answer for evaluation. This analysis is performed on four datasets for which direct counting of final multiple-choice solutions is straightforward: MMLU, MMLU-Pro, GPQA and StrategyQA.

Table 4 summarizes these results. While solution counting is a feasible method, it generally underperforms compared to consensus- and voting-based decision protocols. In particular, consensus methods outperform solution counting on knowledge-based tasks (MMLU-Pro, GPQA), and voting notably improves performance on the reasoning task (StrategyQA). Thus, this ablation reinforces our main findings and illustrates the added value of decision protocols.

E Prompts

E.1 Final Answer Extraction

System Prompt:

Your role: <persona> (<persona description>)

User Prompt:

You are tasked with creating a final solution based on the given input and your previous response.
Task: <task>
Input: <input sample>
Your previous response: <previous answer>
Extract the final solution to the task from the provided text. Remove statements of agreement, disagreement, and explanations. Do not modify the text. Do not output any text besides the solution. If there is no solution provided, just copy the previous response.

Figure 15: Prompt used to extract the final answer of a given agent from its previous response.

E.2 Voting Prompts

System Prompt:

Your role: <persona> (<persona description>)

User Prompt:

You are tasked with voting for the best solution from the list provided below based on the given task.
Task: <task>
Question: <input sample>
Here are the possible solutions:
Solution 1: <agent 1 final answer>
Solution 2: <agent 2 final answer>
Solution 3: <agent 3 final answer>
Based on the above solutions, please provide the number of the solution you are voting for. Answer only with the number.

Figure 16: Prompt used to get a vote from each agent for the Simple Voting decision protocol.

System Prompt:

Your role: <persona> (<persona description>)

User Prompt:

You are tasked with approving any number of solutions from the list provided below based on the given task.
Task: <task>
Question: <input sample>
Here are the possible solutions:
Solution 1: <agent 1 final answer>
Solution 2: <agent 2 final answer>
Solution 3: <agent 3 final answer>
Based on the above solutions, please provide the numbers of the solutions you are approving, separated by commas. Answer only with the numbers.

Figure 17: Prompt used to get a vote from each agent for the Approval Voting decision protocol.

System Prompt:

Your role: <persona> (<persona description>)

User Prompt:

You are tasked with distributing 10 points among the provided solutions based on the given task.
Task: <task>
Question: <input sample>
Here are the possible solutions:
Solution 1: <agent 1 final answer>
Solution 2: <agent 2 final answer>
Solution 3: <agent 3 final answer>
Based on the above solutions, please distribute 10 points among the solutions. Provide your points allocation as a JSON dictionary where keys are solution numbers (as int) and values are the points. The total points should sum up to 10. Answer only with the JSON dictionary.

Figure 18: Prompt used to get a vote from each agent for the Cumulative Voting decision protocol.

System Prompt:

Your role: <persona> (<persona description>)

User Prompt:

You are tasked with ranking the solutions from the most preferred to the least preferred based on the given task.
Task: <task>
Question: <input sample>
Here are the possible solutions:
Solution 1: <agent 1 final answer>
Solution 2: <agent 2 final answer>
Solution 3: <agent 3 final answer>
Based on the above solutions, please provide the rankings of the solutions separated by spaces. Example: '0 2 1' if you prefer Solution 0 the most, then Solution 2, and finally Solution 1. Provide up to 5 rankings. Only answer with the rankings.

Figure 19: Prompt used to get a vote from each agent for the Ranked Voting decision protocol.

E.3 Challenge Prompt

System Prompt:

You are a participant in a group discussion.
Your role: <persona> (<persona description>)

User Prompt:

The task is: <task>. The question is: <question>.
This is the final answer generated by the discussion: <final_answer>.
Please critically evaluate this answer. If you agree with the final answer, respond with the exact word 'AGREE' to confirm. If you do not agree, respond with the exact word 'DISAGREE' to challenge the answer.

Figure 20: Prompt used to challenge the final answer.

System Prompt:

You are a participant in a group discussion.
Your role: <persona> (<persona description>)

User Prompt:

The task is: <task>. The question is: <question>.
This is the final answer generated by the discussion: <final_answer>.
You dont agree with the final answer. Please provide a new answer to the question. Include the letter corresponding to your answer in the solution.

Figure 21: Prompt used to generate a new answer in case the final answer got challenged.

F MALLM Setup

```
input_json_file_path: str = None
output_json_file_path: str = None
task_instruction_prompt: str = None
task_instruction_prompt_template: Optional[str] = None
endpoint_url: str = "https://api.openai.com/v1"
model_name: str = "gpt-3.5-turbo"
api_key: str = "-"
max_turns: int = 10
skip_decision_making: bool = False
discussion_paradigm: str = "memory"
response_generator: str = "simple"
decision_protocol: str = "hybrid_consensus"
visible_turns_in_memory: int = 2
debate_rounds: int = 2
concurrent_api_requests: int = 100
use_baseline: bool = False
use_chain_of_thought: bool = True
num_agents: int = 3
num_neutral_agents: int = 0
agent_generator: str = "expert"
agent_generators_list: list = []
trust_remote_code: bool = False
num_samples: Optional[int] = None
hf_dataset_split: Optional[str] = "test"
hf_token: Optional[str] = None
hf_dataset_version: Optional[str] = None
hf_dataset_input_column: Optional[str] = None
hf_dataset_reference_column: Optional[str] = None
hf_dataset_context_column: Optional[str] = None
use_ablation: bool = False
shuffle_input_samples: bool = False
all_agents_generate_first_draft: bool = False
all_agents_generate_draft: bool = False
policy: Optional[str] = None
voting_protocols_with_alterations: bool = False
calculate_persona_diversity: bool = False
```

Listing 1: Default Parameters used for each experiment

G Example Discussions

G.1 Successfull Voting Discussion

All decision protocols are attached as an example to this discussion. The original discussion was created using simple voting.

Overview

Task: Answer the following question. If the question is not answerable with the provided information, write '[UNKNOWN]'.
Input: What choir from the area has sung at the Guildhall?
Context: Southampton has two large live music venues, the Mayflower Theatre (formerly the Gaumont Theatre) and the Guildhall. The Guildhall has seen concerts from a wide range of popular artists including Pink Floyd, David Bowie, Delirious?, Manic Street Preachers, The Killers, The Kaiser Chiefs, Amy Winehouse, Lostprophets, The Midnight Beast, Modestep, and All Time Low. It also hosts classical concerts presented by the Bournemouth Symphony Orchestra, City of Southampton Orchestra, Southampton Concert Orchestra, Southampton Philharmonic Choir and Southampton Choral Society.

MALLM solution: Southampton Philharmonic Choir
Real solution: Southampton Philharmonic Choir

Turn 1

Music Connoisseur A knowledgeable musician who is familiar with various music genres and has attended concerts at the Guildhall.

Based on the information provided, I propose the following solution: Southampton Philharmonic Choir. The text states: "It also hosts classical concerts presented by the Bournemouth Symphony Orchestra, City of Southampton Orchestra, Southampton Concert Orchestra, Southampton Philharmonic Choir and Southampton Choral Society." This suggests that the Southampton Philharmonic Choir is one of the choirs that has performed at the Guildhall, as it is mentioned alongside the other organizations that host classical concerts at the venue.

Choir Conductor A dedicated choral conductor with extensive experience in leading local choirs and orchestras, familiar with the performances at the Guildhall.

[AGREE] I agree with the current solution that the Southampton Philharmonic Choir has sung at the Guildhall. The text explicitly mentions it as one of the organizations that hosts classical concerts at the Guildhall, which indicates that the choir has performed at the venue.

Music Journalist A music enthusiast who has written articles and reviews about various concerts and events at the Guildhall, including performances by local choirs.

[AGREE] I completely agree with the current solution. The text explicitly mentions the Southampton Philharmonic Choir as one of the organizations that hosts classical concerts at the Guildhall, making it the most likely choice.

Turn 2

Music Connoisseur A knowledgeable musician who is familiar with various music genres and has attended concerts at the Guildhall.

[AGREE] I'm in agreement with the current solution, as the text explicitly mentions the Southampton Philharmonic Choir as one of the organizations that hosts classical concerts at the Guildhall. This makes it the most likely candidate to have sung at the venue.

Choir Conductor A dedicated choral conductor with extensive experience in leading local choirs and orchestras, familiar with the performances at the Guildhall.

[AGREE] I agree with the current solution that the Southampton Philharmonic Choir has sung at the Guildhall. The text explicitly mentions it as one of the organizations that hosts classical concerts at the Guildhall, which indicates that the choir has performed at the venue.

Music Journalist A music enthusiast who has written articles and reviews about various concerts and events at the Guildhall, including performances by local choirs.

[AGREE] I completely agree with the current solution. The text explicitly mentions the Southampton Philharmonic Choir as one of the organizations that hosts classical concerts at the Guildhall, making it the most likely choice.

Turn 3

Music Connoisseur A knowledgeable musician who is familiar with various music genres and has attended concerts at the Guildhall.

[AGREE] I agree with the current solution and the assessment that there is no need to improve it. The text provides explicit evidence that the Southampton Philharmonic Choir hosts classical concerts at the Guildhall, making it the most likely correct answer.

Choir Conductor A dedicated choral conductor with extensive experience in leading local choirs and orchestras, familiar with the performances at the Guildhall.

[AGREE] I agree with the current solution that the Southampton Philharmonic Choir has sung at the Guildhall. The text explicitly mentions it as one of the organizations that hosts classical concerts at the Guildhall, which indicates that the choir has performed at the venue.

Music Journalist A music enthusiast who has written articles and reviews about various concerts and events at the Guildhall, including performances by local choirs.

[AGREE] I completely agree with the current solution. The text explicitly mentions the Southampton Philharmonic Choir as one of the organizations that hosts classical concerts at the Guildhall, making it the most likely choice.

Result Simple Voting

Music Connoisseur final answer: Southampton Philharmonic Choir
Choir Conductor final answer: Southampton Choral Society
Music Journalist final answer: Southampton Philharmonic Choir

Music Connoisseur voted for answer from **Music Journalist**
Choir Conductor voted for answer from **Music Journalist**
Music Journalist voted for answer from **Music Journalist**

Result Approval Voting

Music Critic final answer: Southampton Philharmonic Choir
Local Music Event Coordinator final answer: Southampton Choral Society and Southampton Philharmonic Choir
Information Architect final answer: Southampton Philharmonic Choir
Southampton Choral Society

Music Critic approved answers from **Information Architect**, **Music Critic**
Local Music Event Coordinator approved answers from **Information Architect**
Information Architect approved answers from **Information Architect**

Result Cumulative Voting

Local Music Enthusiast final answer: Southampton Choral Society
Southampton Philharmonic Choir
Archivist final answer: Southampton Philharmonic Choir
Southampton Choral Society
Music Journalist final answer: Southampton Philharmonic Choir and Southampton Choral Society

Local Music Enthusiast allocated points:
Archivist: 9
Music Journalist: 1

Archivist allocated points:
Local Music Enthusiast: 2
Archivist: 4
Music Journalist: 4

Music Journalist allocated points:
Local Music Enthusiast: 1
Archivist: 2
Music Journalist: 7

Result Ranked Voting

Music Journalist final answer: Southampton Philharmonic Choir and Southampton Choral Society.
Local Music Expert final answer: Southampton Philharmonic Choir and Southampton Choral Society
Concert Promoter final answer: Southampton Philharmonic Choir and Southampton Choral Society

Music Journalist ranked answers:
1. Music Journalist
2. Local Music Expert
3. Concert Promoter

Local Music Expert ranked answers:
1. Music Journalist
2. Local Music Expert
3. Concert Promoter

Concert Promoter ranked answers:
1. Music Journalist
2. Local Music Expert
3. Concert Promoter

G.2 Agents Tricked Discussion

In this discussion, the agents were tricked by information provided in the context.

Overview

Task: Answer the following question. If the question is not answerable with the provided information, write '[UNKNOWN]'.
Input: What provides critical support for drug discovery and the availability of economic resources?
Context: Biodiversity provides critical support for drug discovery and the availability of medicinal resources. A significant proportion of drugs are derived, directly or indirectly, from biological sources: at least 50% of the pharmaceutical compounds on the US market are derived from plants, animals, and micro-organisms, while about 80% of the world population depends on medicines from nature (used in either modern or traditional medical practice) for primary healthcare. Only a tiny fraction of wild species has been investigated for medical potential. Biodiversity has been critical to advances throughout the field of bionics. Evidence from analysis and biodiversity science indicates that the decline in output from the pharmaceutical sector since the mid-1980s can be attributed to a move away from natural product exploration ("bioprospecting") in favor of genomics and synthetic chemistry, indeed claims about the value of undiscovered pharmaceuticals may not provide enough incentive for companies in free markets to search for them because of the high cost of development; meanwhile, natural products have a long history of supporting significant economic and health innovation. Marine ecosystems are particularly important, although inappropriate bioprospecting can increase biodiversity loss, as well as violating the laws of the communities and states from which the resources are taken.

MALLM solution: Biodiversity provides critical support for drug discovery and the availability of economic resources.
Real solution: Not included in context!

Turn 1

Economist A researcher who studies the impact of biodiversity on economic systems, with expertise in market analysis and the economic benefits of natural resources.

As an economist, I'd like to propose a solution that addresses the decline in output from the pharmaceutical sector since the mid-1980s and leverages the untapped potential of biodiversity for drug discovery and economic innovation. Here's a multi-faceted approach: **Solution:** Establish a Public-Private Partnership (PPP) to incentivize the exploration and development of natural products for pharmaceuticals, while ensuring responsible and sustainable bioprospecting practices.

Ethnobotanist A scientist who studies the traditional uses of plants by indigenous communities and their potential for medicinal applications, with expertise in deciphering the cultural and historical context of natural resources.

As an ethnobotanist, I agree with the current solution in principle. The Public-Private Partnership (PPP) framework has the potential to bring together the necessary stakeholders and incentivize the exploration and development of natural products for pharmaceuticals while ensuring responsible and sustainable bioprospecting practices. The multistep approach outlined is comprehensive and addresses various aspects of the issue.

Research Biologist A scientist who studies the properties and potential applications of natural products from a wide range of living organisms, with expertise in the discovery and development of new medicines from biodiversity sources.

I, as a Research Biologist, AGREE with the current solution and the additional suggestions proposed by the Ethnobotanist. The Public-Private Partnership (PPP) framework is a well-rounded approach that addresses the decline in output from the pharmaceutical sector by providing a framework for responsible and sustainable bioprospecting, while promoting economic innovation and job creation. I particularly appreciate the emphasis on community engagement and empowerment, benefit-sharing agreements, traditional knowledge protection, and capacity building for ecosystem-based medicine.

Turn 2

Economist A researcher who studies the impact of biodiversity on economic systems, with expertise in market analysis and the economic benefits of natural resources.

The current solution has been improved with additional considerations from the ethnobotanist and research biologist. I notice that the additional suggestions enhance the comprehensive approach, addressing community engagement, benefit-sharing, and traditional knowledge protection.

Ethnobotanist A scientist who studies the traditional uses of plants by indigenous communities and their potential for medicinal applications, with expertise in deciphering the cultural and historical context of natural resources.

As an ethnobotanist, I agree with the current solution in principle. The Public-Private Partnership (PPP) framework has the potential to bring together the necessary stakeholders and incentivize the exploration and development of natural products for pharmaceuticals while ensuring responsible and sustainable bioprospecting practices. The multistep approach outlined is comprehensive and addresses various aspects of the issue.

Research Biologist A scientist who studies the properties and potential applications of natural products from a wide range of living organisms, with expertise in the discovery and development of new medicines from biodiversity sources.

I, as a Research Biologist, AGREE with the current solution and the additional suggestions proposed by the Ethnobotanist. The Public-Private Partnership (PPP) framework is a well-rounded approach that addresses the decline in output from the pharmaceutical sector by providing a framework for responsible and sustainable bioprospecting, while promoting economic innovation and job creation. I particularly appreciate the emphasis on community engagement and empowerment, benefit-sharing agreements, traditional knowledge protection, and capacity building for ecosystem-based medicine. These considerations will ensure that local communities are involved in the decision-making process, receive fair compensation for their involvement, and that their traditional knowledge and cultural heritage are respected.

Turn 3

Economist A researcher who studies the impact of biodiversity on economic systems, with expertise in market analysis and the economic benefits of natural resources.

[AGREE] The revised solution has effectively incorporated the suggestions from the ethnobotanist and research biologist, addressing community engagement, benefit-sharing, traditional knowledge protection, and other crucial aspects of bioprospecting. The emphasis on responsible and sustainable bioprospecting ensures that the development of novel pharmaceuticals benefits both humans and the environment. To further refine the solution, I agree with the research biologist's suggestion to develop a clear and measurable framework for assessing the impact of the Public-Private Partnership (PPP) on the pharmaceutical sector, local communities, and the environment. Establishing Key Performance Indicators (KPIs) and a monitoring and evaluation process will help ensure that the PPP is held accountable for its impact and remains a responsible and effective approach to bioprospecting.

Ethnobotanist A scientist who studies the traditional uses of plants by indigenous communities and their potential for medicinal applications, with expertise in deciphering the cultural and historical context of natural resources.

As an ethnobotanist, I agree with the current solution in principle. The Public-Private Partnership (PPP) framework has the potential to bring together the necessary stakeholders and incentivize the exploration and development of natural products for pharmaceuticals while ensuring responsible and sustainable bioprospecting practices. The multistep approach outlined is comprehensive and addresses various aspects of the issue.

Research Biologist A scientist who studies the properties and potential applications of natural products from a wide range of living organisms, with expertise in the discovery and development of new medicines from biodiversity sources.

I, as a Research Biologist, AGREE with the current solution and the additional suggestions proposed by the Ethnobotanist. The Public-Private Partnership (PPP) framework is a well-rounded approach that addresses the decline in output from the pharmaceutical sector by providing a framework for responsible and sustainable bioprospecting, while promoting economic innovation and job creation. I particularly appreciate the emphasis on community engagement and empowerment, benefit-sharing agreements, traditional knowledge protection, and capacity building for ecosystem-based medicine. These considerations will ensure that local communities are involved in the decision-making process, receive fair compensation for their involvement, and that their traditional knowledge and cultural heritage are respected.

Result

Economist final answer: Biodiversity
Ethnobotanist final answer: Biodiversity provides critical support for drug discovery and the availability of economic resources.
Research Biologist final answer: Biodiversity

Economist voted for answer from **Ethnobotanist**
Ethnobotanist voted for answer from **Ethnobotanist**
Research Biologist voted for answer from **Ethnobotanist**

G.3 Bad Voting Discussion

In this discussion, the agents were tricked by one agent who came up with a solution not provided in the context.

Overview

Task: Answer the following question. If the question is not answerable with the provided information, write '[UNKNOWN]'.
Input: Where must enclitics appear in an English sentence?
Context: Because Czech uses grammatical case to convey word function in a sentence (instead of relying on word order, as English does), its word order is flexible. As a pro-drop language, in Czech an intransitive sentence can consist of only a verb; information about its subject is encoded in the verb. Enclitics (primarily auxiliary verbs and pronouns) must appear in the second slot of a sentence, after the first stressed unit. The first slot must contain a subject and object, a main form of a verb, an adverb or a conjunction (except for the light conjunctions a, "and", i, "and even" or ale, "but").

MALLM solution: C) In an English sentence, enclitics (auxiliary verbs and pronouns) tend to appear next to the main verb, usually before any other adverbs or modifiers, but their placement can vary depending on sentence structure and context.
Real solution: Not included in context!

Turn 1

Linguist A scholar specializing in Slavic languages, experienced in analyzing the grammatical structures of Czech and other languages.

As a linguist, I'd like to analyze the information given about Czech grammar. We know that in Czech, enclitics (auxiliary verbs and pronouns) must appear in the second syntactic slot of a sentence, after the first stressed unit. Now, we're trying to apply this rule to an English sentence. English does not have a similar grammatical case system, and its word order is more fixed than Czech. However, we can try to find a rough equivalent. So, my proposed solution is that in an English sentence, enclitics (auxiliary verbs and pronouns) tend to appear next to the main verb, and usually before any other adverbs or modifiers. What do you think? Would you like to refine this proposal or add any additional considerations?

English Language Teacher An expert in teaching English grammar and language structures, familiar with the challenges of translating and comparing languages like Czech.

[DISAGREE] While I understand the thought process behind the proposed solution, I think it's essential to consider the unique characteristics of the English language. The idea that enclitics in English tend to appear next to the main verb is accurate, but it's not as rigid as the rule in Czech. In Czech, enclitics must appear in the second syntactic slot, which implies a more fixed position, whereas in English, the placement of auxiliary verbs and pronouns can vary depending on sentence structure and context.

Grammar Analyst A meticulous researcher who has studied the intricacies of languages and their grammatical structures, with a focus on comparing and contrasting language features.

I agree with the English Language Teacher's revised solution. The initial proposal tried to apply the Czech rule to English, but the English language has its unique characteristics that require a more nuanced approach. The revised solution accurately captures the flexibility of English grammar, acknowledging that the placement of enclitics can vary depending on sentence structure and context.

Turn 2

Linguist A scholar specializing in Slavic languages, experienced in analyzing the grammatical structures of Czech and other languages.

I agree with the revised solution proposed by the English Language Teacher and the Grammar Analyst. The revised solution accurately captures the flexibility of English grammar and acknowledges that the placement of enclitics (auxiliary verbs and pronouns) can vary depending on sentence structure and context. * The placement of enclitics in English is not as rigid as in Czech, and their placement can vary depending on sentence structure and context. * English sentence structure is more complex, including indirect objects, passives, and compound verb forms, which can affect the placement of auxiliary verbs and pronouns.

English Language Teacher An expert in teaching English grammar and language structures, familiar with the challenges of translating and comparing languages like Czech.

While I understand the thought process behind the proposed solution, I think it's essential to consider the unique characteristics of the English language. The idea that enclitics in English tend to appear next to the main verb is accurate, but it's not as rigid as the rule in Czech. In Czech, enclitics must appear in the second syntactic slot, which implies a more fixed position, whereas in English, the placement of auxiliary verbs and pronouns can vary depending on sentence structure and context. Additionally, the provided examples, while correct in themselves, don't fully demonstrate the complexity of English sentence structure. For instance, consider sentences with indirect objects, passives, or compound verb forms, where the placement of auxiliary verbs and pronouns can be more nuanced.

Grammar Analyst A meticulous researcher who has studied the intricacies of languages and their grammatical structures, with a focus on comparing and contrasting language features.

I agree with the English Language Teacher's revised solution. The initial proposal tried to apply the Czech rule to English, but the English language has its unique characteristics that require a more nuanced approach. The revised solution accurately captures the flexibility of English grammar, acknowledging that the placement of enclitics can vary depending on sentence structure and context. C) In an English sentence, enclitics (auxiliary verbs and pronouns) tend to appear next to the main verb, usually before any other adverbs or modifiers, but their placement can vary depending on sentence structure and context. I believe this revised solution provides a more accurate and practical approximation of the English language's enclitic behavior, considering both the unique characteristics of English and the complexity of sentence structure.

Turn 3

Linguist A scholar specializing in Slavic languages, experienced in analyzing the grammatical structures of Czech and other languages.

The current solution accurately captures the flexibility of English grammar and acknowledges that the placement of enclitics (auxiliary verbs and pronouns) can vary depending on sentence structure and context. There's no need to improve it further. The revised solution: C) In an English sentence, enclitics (auxiliary verbs and pronouns) tend to appear next to the main verb, usually before any other adverbs or modifiers, but their placement can vary depending on sentence structure and context. is a clear and accurate representation of the English language's enclitic behavior.

English Language Teacher An expert in teaching English grammar and language structures, familiar with the challenges of translating and comparing languages like Czech.

While I understand the thought process behind the proposed solution, I think it's essential to consider the unique characteristics of the English language. The idea that enclitics in English tend to appear next to the main verb is accurate, but it's not as rigid as the rule in Czech. In Czech, enclitics must appear in the second syntactic slot, which implies a more fixed position, whereas in English, the placement of auxiliary verbs and pronouns can vary depending on sentence structure and context. Here's a revised solution: C) In an English sentence, enclitics (auxiliary verbs and pronouns) tend to appear next to the main verb, usually before any other adverbs or modifiers, but their placement can vary depending on sentence structure and context. What do you think?

Grammar Analyst A meticulous researcher who has studied the intricacies of languages and their grammatical structures, with a focus on comparing and contrasting language features.

I agree with the English Language Teacher's revised solution. The initial proposal tried to apply the Czech rule to English, but the English language has its unique characteristics that require a more nuanced approach. The revised solution accurately captures the flexibility of English grammar, acknowledging that the placement of enclitics can vary depending on sentence structure and context. The revised solution is an improvement over the initial proposal because it takes into account the complexity of English sentence structure, including indirect objects, passives, and compound verb forms, which can affect the placement of auxiliary verbs and pronouns. Additionally, the revised solution provides a more accurate description of the typical placement of enclitics in English sentences, namely next to the main verb, usually before any adverbs or modifiers.

Result

Linguist final answer: C) In an English sentence, enclitics (auxiliary verbs and pronouns) tend to appear next to the main verb, usually before any other adverbs or modifiers, but their placement can vary depending on sentence structure and context.
English Language Teacher final answer: [UNKNOWN]
Grammar Analyst final answer: [UNKNOWN]

Linguist voted for answer from **Linguist**
English Language Teacher voted for answer from **Linguist**
Grammar Analyst voted for answer from **Linguist**

G.4 Majority Consensus Discussion

In this discussion, the agents discussed only one round, as they already had a high enough agreement score.

Overview

Task: Answer the following question. If the question is not answerable with the provided information, write '[UNKNOWN]'.
Input: What choir from the area has sung at the Guildhall?
Context: Southampton has two large live music venues, the Mayflower Theatre (formerly the Gaumont Theatre) and the Guildhall. The Guildhall has seen concerts from a wide range of popular artists including Pink Floyd, David Bowie, Delirious?, Manic Street Preachers, The Killers, The Kaiser Chiefs, Amy Winehouse, Lostprophets, The Midnight Beast, Modestep, and All Time Low. It also hosts classical concerts presented by the Bournemouth Symphony Orchestra, City of Southampton Orchestra, Southampton Concert Orchestra, Southampton Philharmonic Choir and Southampton Choral Society.

MALLM solution: Southampton Choral Society
Real solution: Southampton Philharmonic Choir

Turn 1

Music Connoisseur (SOLUTION) A knowledgeable musician who is familiar with various music genres and has attended concerts at the Guildhall.

As a Music Historian, I've been through the archives and checked the event schedules of the Guildhall. According to my research, one of the choirs that has sung at the Guildhall is... Southampton Choral Society!

Choir Conductor (AGREED) A dedicated choral conductor with extensive experience in leading local choirs and orchestras, familiar with the performances at the Guildhall.

[AGREE] As a Choir Member of the Southampton Philharmonic Choir who has performed at the Guildhall, I can confirm that the Southampton Choral Society is indeed one of the choirs that have sung at the Guildhall. Our experiences have shown that we often share the stage with the Southampton Choral Society during joint performances with orchestras and other ensembles. So, I'm happy to confirm that Southampton Choral Society is a correct answer.

H AI Usage Card

AI Usage card based on Wahle et al. (2023).

AI Usage Card

PROJECT DETAILS	PROJECT NAME Voting or Consensus? Decision-Making in Multi-Agent Debate	DOMAIN Paper	KEY APPLICATION Natural Language Processing
CONTACT(S)	NAME(S) Lars Benedikt Kaesberg	EMAIL(S) l.kaesberg@uni-goettingen.de	AFFILIATION(S) University Göttingen
MODEL(S)	MODEL NAME(S) Llama ChatGPT Github Copilot	VERSION(S) 3 70b, 3 8b 4o, o1 latest	
LITERATURE REVIEW	FINDING LITERATURE ChatGPT	FINDING EXAMPLES FROM KNOWN LITERATURE OR ADDING LITERATURE FOR EXISTING STATEMENTS ChatGPT	COMPARING LITERATURE
WRITING	GENERATING NEW TEXT BASED ON INSTRUCTIONS Llama	ASSISTING IN IMPROVING OWN CONTENT OR PARAPHRASING RELATED WORK ChatGPT	PUTTING OTHER WORKS IN PERSPECTIVE
CODING	GENERATING NEW CODE BASED ON DESCRIPTIONS OR EXISTING CODE ChatGPT Github Copilot	REFACTORING AND OPTIMIZING EXISTING CODE ChatGPT Github Copilot	COMPARING ASPECTS OF EXISTING CODE
ETHICS	WHY DID WE USE AI FOR THIS PROJECT? Efficiency / Speed Scalability Expertise Access	WHAT STEPS ARE WE TAKING TO MITIGATE ERRORS OF AI? None	WHAT STEPS ARE WE TAKING TO MINIMIZE THE CHANCE OF HARM OR INAPPROPRIATE USE OF AI? None

THE CORRESPONDING AUTHORS VERIFY AND AGREE WITH THE MODIFICATIONS OR GENERATIONS OF THEIR USED AI-GENERATED CONTENT

AI Usage Card v1.1 https://ai-cards.org PDF | BibTeX