

Better Language Reasoning Through Multi-Agent Debate

Rajesh Raju
Mila-AI Quebec
University de Montreal
rajesh.raju@mila.quebec

Venkata Kothapalli
Mila-AI Quebec
University de Montreal
venkata.kothapalli@mila.quebec

Giovanni Belval
University de Montreal
giovanni.belval@umontreal.ca

Abstract

This study evaluates the efficacy of Multi-Agent Debate (MAD) strategies, specifically the "Society of Minds" and "Multi-Persona" approaches, on the accuracy of responses from large language models (LLMs) across Math_QA, MedMCQA, and SciQ datasets. Despite using adversarial learning and iterative refinement to improve answers, these collaborative debate strategies do not consistently outperform single-agent baselines. The success of MAD strategies varies significantly depending on the model's specific characteristics and the dataset used. Our research underscores the need for customized strategies that exploit the unique capabilities of each model to enhance the overall performance of adversarial learning environments and ensure the reliability of LLM.¹

1 Introduction

Multi-agent debate (MAD) is an emerging concept in the field of artificial intelligence (AI) and natural language processing (NLP) where multiple AI agents engage in a structured argument or discussion on a given topic and has revealed the substantial capacity of large language models (LLMs) to improve their performance on reasoning tasks. The idea is to simulate a human-like debate environment where each agent represents a different viewpoint or argument strategy. Multi-agent debate, as opposed to single-agent LLMs, offers a range of advantages, especially in the context of enhancing AI capabilities and providing more nuanced, balanced, and comprehensive outputs. Some of the key advantages include diverse perspectives and counterarguments, enhanced critical thinking and rebuttal skills, dynamic interaction and adaptability, improved decision making and bias detection and mitigation.

In our research, we assess the effectiveness of MAD strategies across various question-answering scenarios. We utilize mathematical problems from the Math_QA dataset (Aida et al., 2019), medical inquiries from the MedMCQA dataset (Ankit et al. 2022), and scientific questions from the SciQ dataset (Welb et al., 2017) as sources for our queries. The experimental setup involves deploying MAD strategies to address a diverse array of questions. Specifically, we implement two distinct MAD approaches: "Society of Minds" (SoM-MAD) and "Multi-Persona" (MP-MAD). These models are tasked with not only providing answers but also articulating the reasoning behind their conclusions. We incorporate a debate mechanism that enables the models to challenge and refine their responses or the supporting arguments based on the outcomes of the debates.

In the domain of related work, MAD strategies have been implemented across various tasks, yet there remains an absence of comprehensive research into the optimal number of agents and debate rounds necessary to maximize the effectiveness of MAD models. Additionally, investigations into whether there is a linear relationship between the number of rounds or agents and the peak performance of these models are sparse, and there is no clear consensus on whether increasing the number of agents or rounds consistently improves performance compared to baseline single-agent models. Our study addresses these gaps by methodically examining these questions. Unlike previous research that primarily utilizes GPT models, our study employs a diverse array of LLMs to conduct an extensive evaluation of MAD strategies across a broad spectrum of models.

2 Related works

The integration of LLMs in autonomous multi-agent systems has garnered substantial interest

¹https://github.com/rajeshkochi444/mad_llm

across both industry and academic sectors (Du et al., 2023; Zhuge et al., 2023; Smit et al., 2023; Liang et al., 2023; Hong et al., 2023; Szot et al., 2023; Pham et al., 2023; Park et al., 2023; Liu et al., 2023; Chan et al., 2023), CIPHER, (Pham et al., 2023). The Generative Agents project (Park et al., 2023) has simulated a microcosm of 25 agents within a "town", enabling the study of language interactions, societal comprehension, and shared cognition. Similarly, the Natural Language-Based Society of Mind (NLSOM) framework (Zhuge et al., 2023) delineates a system where agents with diversified roles engage in serial 'mindstorms' to tackle intricate tasks. Liang et al., (2023) incorporated a "judge" to resolve tiebreakers and determine the final answer. This approach promotes divergent thinking in LLMs and bypasses the feedback bottleneck, demonstrating superior performance across various datasets. CIPHER, (Pham et al., 2023) a communication regime for LLMs, which removes token sampling and allows LLMs to communicate beliefs via raw transformer output embeddings, showcasing superior performance in reasoning tasks compared to traditional natural language MAD methods. MetaGPT (Hong et al., 2023) is a meta-programming framework that integrates SOPs into prompt sequences, streamlining workflows and reducing errors in LLM-based multi-agent systems, surpassing previous chat-based systems in collaborative software engineering benchmarks.



Figure 1: Graphical illustration of the multi-agent debate (MAD) framework.

3 Methods

3.1 "Society of Minds" MAD strategy (SoM-MAD)

The initial phase of our project utilized the "Society of Minds" (SoM-MAD) model, a unique framework designed for collective intelligence and

problem-solving. Key features of this model include:

- **Multiple Agents Collaboration:** Unlike traditional models where a single agent tackles problems independently, the SoM framework integrates the expertise of numerous AI agents. Each agent can specialize in different areas or methodologies, offering diverse perspectives on a problem.
- **Effective Collaborations:** The SoM model encourages effective collaboration among agents. By leveraging the strengths of each agent, the model facilitates a more comprehensive and nuanced problem-solving process.
- **Summarization of Answers:** After each debate round, a mediator summarizes the answers provided by the agents. This summary is then added to the historical context that all agents can access in future rounds. This mechanism ensures that knowledge is accumulated and shared, enhancing the overall decision-making process.
- **Mediator Role:** The mediator plays a critical role by synthesizing the debate outcomes and maintaining the coherence of the discussions. This ensures that the collective intelligence of the agents is effectively harnessed and utilized in solving complex problems.

These features make the SoM model a robust and dynamic framework for harnessing the power of collective AI in solving diverse and complex challenges.

3.2 "Multi-Persona" MAD strategy (MP-MAD)

In the second phase of the project, we implemented a refined version of the MAD strategy known as the Multi-Persona MAD strategy (MP-MAD), which incorporates distinct agent personalities to encourage a breadth of perspectives. This strategy includes:

- **Dual-Persona Setup:** Agents are prompted with specific roles to foster divergent outcomes. One agent assumes the role of an "Affirmative Agent" (often likened to an angel), promoting and supporting ideas, while another takes on the "Negative Agent" role (akin to a devil), challenging and critiquing

proposals. This dynamic ensures a comprehensive evaluation of ideas from both supportive and critical viewpoints.

- **Judge Agent:** Central to this strategy is the Judge Agent, who oversees the debate process. This agent listens to the inputs from both the affirmative and negative agents, then synthesizes these arguments to reach a balanced and informed final solution.
- **Critic Role:** Additionally, our model introduces a Critic, an agent specifically designed to oppose arguments presented by its counterpart, enhancing the rigor of the debate. This Critic does not merely react but provides structured counterarguments to ensure that all proposals are robustly tested before any conclusions are reached.

This structured approach, by promoting a dynamic interplay of affirmations and criticisms, helps in deriving well-rounded and thoroughly vetted solutions, maximizing the efficacy of collective intelligence in our decision-making processes.

3.3 CrewAI

Our implementation utilized CrewAI, an advanced framework built upon Langchain, designed for orchestrating role-playing, autonomous AI agents. CrewAI supports collaborative intelligence, enabling these agents to work together seamlessly to tackle complex tasks. The framework allows each agent to assume specific roles, share objectives, and autonomously delegate tasks within the group, creating a cooperative unit similar to a crew. This setup enhances the ability of the agents to interact dynamically and efficiently, fostering a collaborative environment where they can exchange information and jointly address challenges more effectively.

CrewAI has the following components:

- **Agents:** Individual members of the team, each possessing unique skills.
- **Tasks:** Specific assignments within CrewAI, possibly requiring collaborative efforts from multiple agents.
- **Crew:** Groups of agents in CrewAI working in concert to accomplish collective tasks.

Each component has specific attributes associated with it, which are listed in Tables 1 - 3.

Agent Attribute	Description
Role	The function or position the agent fulfills within the team.
Goal	The objective the agent aims to achieve.
Backstory	The narrative providing context for the agent's behavior and objectives.
LLM	Indicates the inclusion of language processing capabilities.
Tools	Instruments or software the agent can utilize to perform tasks.
Allow Delegation	The ability to delegate tasks or roles to other agents.
Memory	The capacity to retain and recall past experiences or information.

Table 1: Agent Attributes Description

Task Attribute	Description
Description	A detailed explanation of the task.
Agent	The specific agent(s) assigned to the task.
Expected Output	The desired result or product of the task.
Context	Background information relevant to the task.
Goal	The end objective the task is intended to accomplish.
Callback	A mechanism to return the result or status of the task upon completion.

Table 2: Task Attributes Description

Crew Attribute	Description
Tasks	The assignments the crew is responsible for completing.
Agents	The collective of individual agents comprising the crew.
Process	The methods and sequences followed to complete tasks.
Memory	Shared knowledge and experiences retained by the crew.
Call Backs	Procedures in place for reporting progress or results back to a central system or overseer.

Table 3: Crew Attributes Description

3.4 Our implementation for MAD

In our Study of SoM MAD strategy, we developed specific prompts for agents involved in tasks. For a mathematical problem-solving scenario, we structured a debate between two agents labeled "Student" and "Professor." The Student agent began the debate by attempting to solve a mathematical question, providing a detailed explanation of each step and the rationale behind them. Following this, the Professor agent reviewed the Student's work, providing feedback and the correct solutions while also engaging in a discussion. This discussion included proposing alternative methods and challenging the Student's approach, prompting corrections and deeper analysis where necessary.

```

def task_grading(math_question, agent, task_context_solving):
    return Task(
        description=dedent(f"""
            For the solution provided for the math problem: {math_question}, debate on the student's answer.
            You must review the solution, give feedback to student and provide correct solutions.
            If the solution is wrong, give correct solution with explanation.
            Provide alternative pathways or easy methods to solve the math question if any.
            If you found the answer is wrong, ask the student to improve or correct his answer using your comments/feedback
            including alternative pathways as suggested.""",
            ),
        expected_output=dedent(f"""
            A detailed feedback of student's solution. Give alternative pathways to solutions and provide correct solutions.""",
            ),
        context=[task_context_solving],
        agent=agent,
        output_file='grading_task.txt'
    )

```

Figure 2: Prompt for the "Professor" in the CrewAI Task Setup Employing the SoM-MAD Strategy for Math Reasoning

The Student was likewise tasked to actively participate in this discourse, contesting the Professor’s critiques and suggestions, and refining their solutions based on the feedback.

In a parallel approach to our mathematical problem-solving model, we developed a similar framework for debating science and medical related questions from SciQ and MedMCQA datasets. This model involves two agents, designated as Student1 and Student2, each tasked with providing accurate responses to science/medical questions, supported by logical arguments and reasoning. Both Student1 and Student2 are programmed to engage in a structured debate, where each agent presents and defends its answer using scientifically valid arguments. The objective is not only to ascertain the correct answer but also to simulate a realistic scientific discourse, underscoring the importance of evidence-based reasoning in science.

Additionally, we incorporated a Mediator agent within the framework of our study to assess the arguments and solutions advanced by the debating agents for both math and science/medical reasoning. The primary function of the Mediator is to synthesize the inputs from the debate and derive a conclusive solution for the presented mathematical and scientific problems. Throughout each round, the Mediator critically evaluates the arguments, reasoning, and solutions proffered by the agents. This role is pivotal as it involves a thorough scrutiny of the validity and relevance of each argument, with a particular focus on how robustly each agent substantiates its response using scientific and mathematical reasoning.

Immediately following the initiation of the debate by the Student agent, the Mediator performs an initial evaluation of the Student’s arguments and reasoning. This preliminary analysis serves as the Single Agent solution, establishing a baseline for

subsequent comparative assessments in the study.

Figure 2 illustrates the setup design for the CrewAI agent and tasks within the MAD framework, specifically tailored for mathematical reasoning tasks. For information on setups designed for science and medical reasoning tasks, please refer to the corresponding GitHub repository. Additionally, we have included examples of prompts for CrewAI tasks along with the outputs from the MAD strategies in the appendix Figures 9 - 14.

4 Experiments

4.1 Datasets

In our Multi-Agent Debate (MAD) study, we leveraged the Math_QA dataset (Aida et al., 2019), which is aimed at mathematical problem-solving. This dataset contains a range of problem statements, each paired with a quintet of potential answers, including the correct one. Our methodology, however, diverged from directly predicting the correct answer choice. Instead, we engaged the model in generating the correct numerical value as its answer, enabling us to evaluate the model’s precision by comparing its output against the correct answers enumerated within the dataset.

We also utilized the SciQ dataset (Welbl et al., 2017), which consists of 13,679 science examination questions sourced from crowdsourcing. These questions span a variety of subjects, such as Physics, Chemistry, and Biology, and are structured in a multiple-choice format, each with four options. The aim here was for the model to discern and select the most accurate answer following a structured debate.

Furthermore, the MedMCQA dataset (Ankit et al. 2022), served as an expansive and relevant tool for assessing medical examination questions as they are framed in the real world. With close to 194,000 questions, its extensive breadth is designed

to mirror the intricate nature of actual exam settings. While the dataset mainly features multiple-choice questions with several correct answers, our focus was narrowed down to those questions with a singular correct option.

For each of these datasets, due to computational limits, we strategically selected a subset of 100 examples. This was a necessary compromise to ensure the feasibility of our study within the bounds of our resource capabilities.

4.2 Baselines

Our study employed datasets distinct from those reported in existing research. To provide a meaningful context for our analysis, we used the performance of GPT-3.5-Turbo as our baseline. This approach aligns with our objective of evaluating the MAD model’s efficacy across various tasks. To establish a clear benchmark, we conducted evaluations using a single-agent model for all tasks. In subsequent research, for each LLM under consideration, we will adopt the performance of its respective single-agent model as the baseline. This approach ensures a consistent and relevant benchmark for comparative analysis across different LLMs.

4.3 Evaluation Methods

In the evaluation of the Math_QA dataset, our methodology employed numeric accuracy as the primary metric. A score of 1 was assigned to the model’s response if it correctly matched the numeric answer associated with the correct label in the dataset, within a margin of 0.02; otherwise a score of zero will be awarded.

For SciQ and MedMCQA datasets comprising Single-Correct Multiple Choice Questions (MCQs), accuracy was the chosen metric. Here, a score of 1 was awarded when the model’s response accurately aligned with the correct option provided in the dataset.

In both scenarios, the model’s overall performance was assessed based on the cumulative total of these scores.

4.4 Experimental Results

In our experiments on the "Society of Minds" and "Multi-Persona" MAD strategies, we employ the following LLMs (Table 4). For all experiments, we employ consistent hyperparameter settings to facilitate direct comparison. These settings are given in Table 5.

Table 4: List of Models Used in the Experiments

Model Name
LLama2-7B-Chat
Mistral-7B-Instruct-v0-2
Mixtral-8x7B-Instruct-v0.1
Vicuna-7B-v1.5
Zephyr-7b-beta
Solar-10.7B-Instruct-v1.0
GPT3.5-Turbo
WizardMath-7V-V1.1

Table 5: Hyperparameters

Hyperparameter	Value
Max_new_tokens	1024
Top_k	10
Top_p	0.95
Temperature	0.7

4.5 Results and Analysis

In Figure 3, we present the performance outcomes of various LLMs employing the Society of Minds (SoM) MAD strategy across multiple debate rounds on the Math_QA dataset. The depicted bar chart categorizes the performance of several models, namely LLama2-7B-Chat, Mistral-7B-Instruct-v0.2, Mixtral-8x7B-Instruct-v0.1, Vicuna-7B-v1.5, Zephyr-7b-beta, WizardMath-7B-V1.1, Solar-10.7B-Instruct-v1.0, and GPT3.5-Turbo. The x-axis lists these models, while the y-axis quantifies the percentage of correct responses achieved by each.

The baseline performance in Round 0 (single-agent scenario) indicates notably low accuracies for LLama2-7B-Chat, Mistral-7B-Instruct-v0.2, Vicuna-7B-v1.5, and Zephyr-7b-beta. Conversely, Mixtral-8x7B-Instruct-v0.1, WizardMath-7B-V1.1, and Solar-10.7B-Instruct-v1.0 demonstrate moderate accuracy’s of 32%, 21%, and 18% respectively. GPT3.5-Turbo outperforms all other models at this stage with a baseline accuracy of 46%.

During subsequent debate rounds, the models exhibit varied trajectories of performance. Mixtral-8x7B-Instruct-v0.1 initially improves to 35% in the first debate round but subsequently declines to 32% in the second round, ending with a further drop to 24% in the third round. Solar-10.7B-Instruct-v1.0 sees an increase in the second round but falls to lower levels in the final round. Notably, GPT3.5-Turbo is the only model that consistently

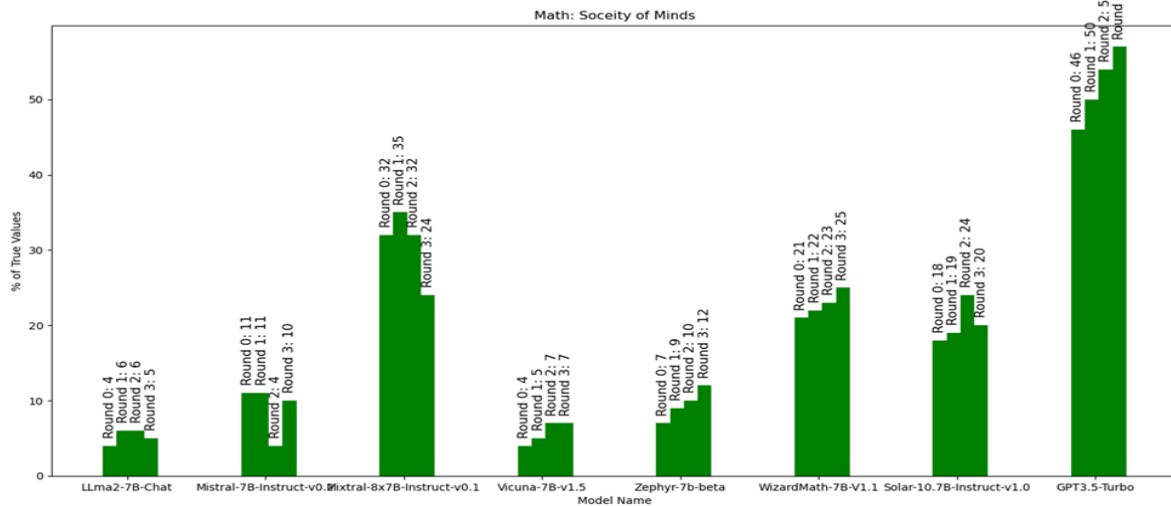


Figure 3: Performance of the SoM-MAD strategy in different debate rounds for various LLMs using the Math_Q dataset, where round 0 represents the baseline single-agent performance.

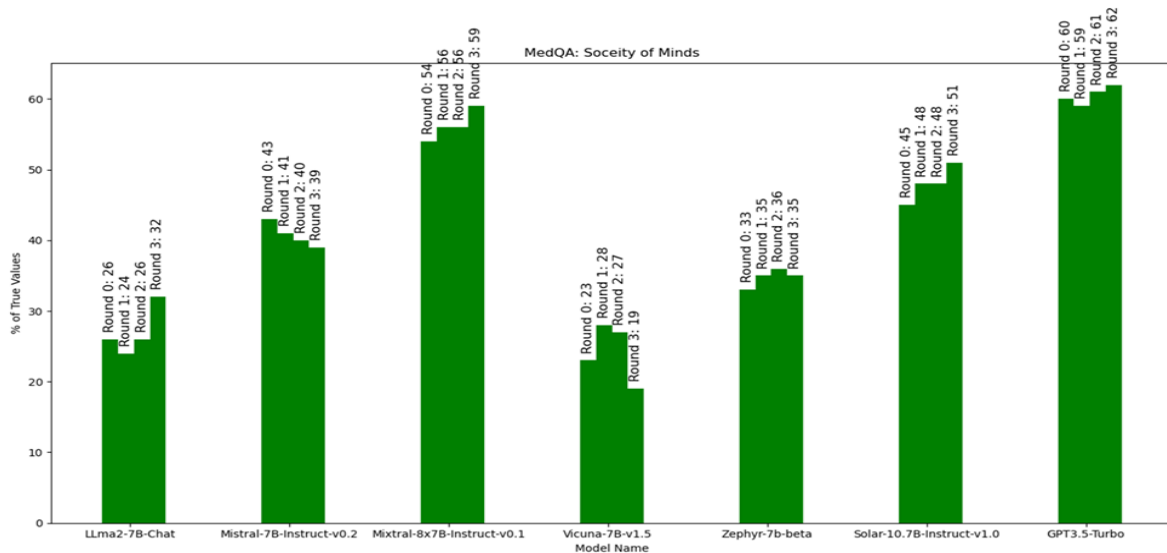


Figure 4: Performance of the SoM-MAD strategy in different debate rounds for various LLMs using the MedMCQA dataset, where round 0 represents the baseline single-agent performance.

improves across all debate rounds, achieving an overall increase of 11% from its baseline by the end of Round 3. This evidence suggests that while the MAD strategy can facilitate incremental learning and improvement in model accuracy, the extent of this benefit varies significantly across different models, with GPT3.5-Turbo showing the most pronounced and consistent gains.

Figures 3 - 5 presents a detailed analysis of various large language models (LLMs) across the MedMCQA and SciQ datasets using a Multi-Agent Debate (MAD) strategy. When comparing these results to those obtained from the

Math_QA dataset, it is evident that all models achieved better performance in predicting correct answers for science and medical-related questions. Four models—Mistral-7B-Instruct-v0.2, Mixtral-8x7B-Instruct-v0.1, Solar-10.7B-Instruct-v1.0, and GPT3.5-Turbo—demonstrate notably superior performance on both the SciQ and MedMCQA datasets. Specifically, GPT3.5-Turbo and Mixtral-8x7B-Instruct-v0.1 stand out, ranking first and second respectively in performance for single-agent baselines. These results underscore the effectiveness of these models in handling complex queries within the specialized domains of science

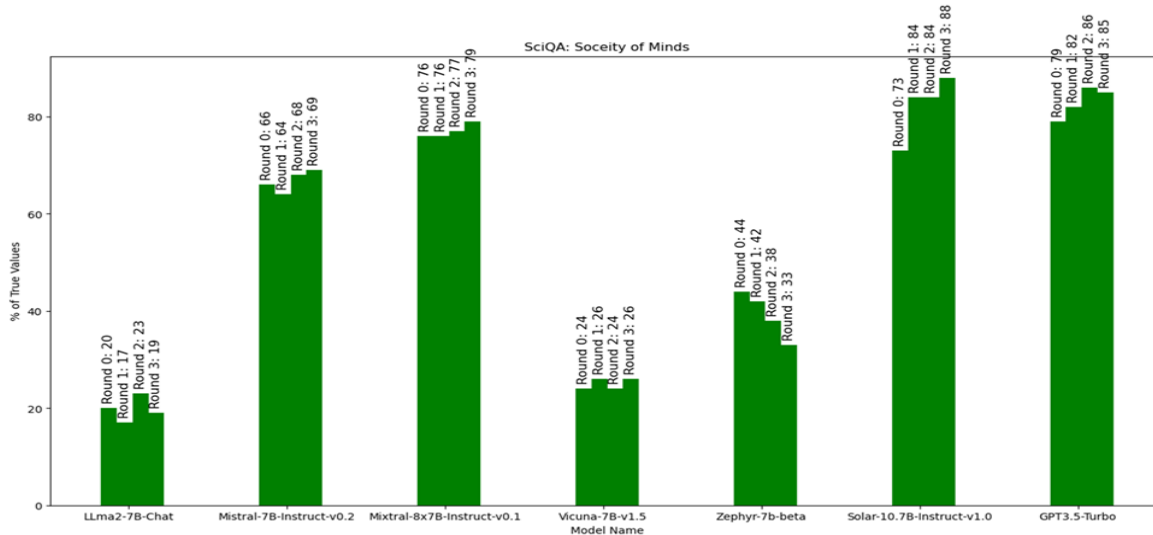


Figure 5: Performance of the SOM-MAD strategy in different debate rounds for various LLMs using the SciQ dataset, where round 0 represents the baseline single-agent performance.

and medicine, suggesting that their architectures or training methodologies are particularly well-suited to benefit from the iterative and adversarial nature of the MAD strategy.

The analysis on MedMCQA dataset indicates that multiple rounds of debate do not consistently enhance model performance. Notably, certain models such as Mistral-7B-Instruct-v0.2 and Vicuna-7B-v1.5 exhibit a decline in performance throughout the debate process, while others show only marginal improvements. Specifically, Mistral-7B-Instruct-v0.2 shows a consistent decrease in performance, falling from 43% accuracy in the initial round to 39% by the conclusion. Vicuna-7B-v1.5 initially shows a 5% improvement over the baseline in the first round but then experiences a decline in the subsequent rounds. In contrast, Llama2-7B-Chat demonstrates an initial performance decrement in the first round but subsequently improves, reaching a 6% increase over the baseline by the third round. Both Mixtral-8x7B-Instruct-v0.1 and Solar-10.7B-Instruct-v1.0 achieve modest gains of 5% and 6%, respectively, in the final round compared to the baseline. However, GPT3.5-Turbo does not exhibit significant performance enhancements SoM-MAD strategy, indicating that the effectiveness of this iterative and adversarial debate format varies significantly across different models.

Further analysis on the SciQ dataset corroborates observations similar to those found in the Math_QA and MedMCQA datasets. Initial performance metrics reveal that Llama2-7B-Chat, Vicuna-7B-v1.5,

and Zephyr-7b-beta start with baseline accuracies of 20%, 24%, and 44%, respectively. In comparison, other models demonstrate higher initial performance: Mistral-7B-Instruct-v0.2 at 66%, Mixtral-8x7B-Instruct-v0.1 at 76%, Solar-10.7B-Instruct-v1.0 at 73%, and GPT3.5-Turbo at 79%. However, the application of the SoM-MAD strategy does not result in significant improvements in performance for most models.

Zephyr-7b-beta consistently shows a decline in performance across subsequent rounds. In contrast, models that exhibited relatively higher initial accuracies such as Mistral-7B-Instruct-v0.2 and Mixtral-8x7B-Instruct-v0.1 only show a marginal improvement of 3%, while GPT3.5-Turbo improves by 6%, increasing its accuracy from 79% to 85%. Notably, Solar-10.7B-Instruct-v1.0 demonstrates a substantial improvement in the final round, with accuracy jumping from 73% at baseline to 88% in the third round of debate. These results suggest significant variability in the effectiveness of the SoM-MAD strategy across different models and datasets, highlighting the influence of individual model characteristics and capabilities on performance outcomes in debate settings.

Figures 6 - 8 in the appendix present the outcomes of employing the Multi-Persona Multi-Agent Debate (MP-MAD) strategy across the Math_QA, MedMCQA, and SciQ datasets. This strategy yielded performance patterns for all models that were generally similar to those observed with the Society of Mind (SoM) of Multi-Agent De-

bate (SoM-MAD) strategy. With the exception of the GPT3.5-Turbo model, which demonstrated an improvement of 11%—mirroring its performance under the SoM-MAD strategy—most models did not show significant performance enhancements through the multi-agent debate process.

For the MedMCQA dataset, a consistent performance decline was noted for models such as Mistral-7B-Instruct-v0.2, Mixtral-8x7B-Instruct-v0.1, and Zephyr-7b-beta. Conversely, Solar-10.7B-Instruct-v1.0 and GPT3.5-Turbo, which had better baseline performances, exhibited only nominal improvements of 4% and 3% respectively in accuracy as the debates progressed.

Similarly, for the SciQ dataset under the MP-MAD strategy, certain models like Mistral-7B-Instruct-v0.2, Mixtral-8x7B-Instruct-v0.1, and GPT3.5-Turbo showed a consistent decline in performance. Specifically, Mistral-7B-Instruct-v0.2 exhibited a notable reduction of 16% to 18% from the baseline to the third round of debate. GPT3.5-Turbo experienced a 4% drop in the final round, with an interim drop from an 84% baseline to 71% in the second round. In contrast, Solar-10.7B-Instruct-v1.0 showed a consistent improvement, rising from a 77% baseline to 87% in the final debate round. These findings highlight the complex dynamics and varied efficacy of the MP-MAD strategy across different models and datasets.

In scientific evaluations of large language models (LLMs) using debate-based learning strategies, one critical factor contributing to the inconsistent performance and, at times, the decline in model accuracy is the potential confusion induced by the debate process itself. Detailed analysis indicates that the models may diverge from correct responses during the iterative steps of the debate. This divergence often stems from the mediator agent’s inability to definitively determine the most accurate answer after each debate round. Consequently, this indecision can lead to situations where initially correct answers are inadvertently swapped for incorrect ones. Such outcomes suggest that the debate format, while intended to refine model outputs through adversarial interaction, might instead lead to increased uncertainty and error propagation if the mediator agent lacks robust mechanisms to reliably identify and endorse the most accurate conclusions drawn from the debate rounds. This highlights a significant challenge in the implementation of multi-agent debate strategies, emphasizing the need for improved design and training of medi-

ator agents to enhance decision-making accuracy in complex debate scenarios.

5 Conclusion

The comprehensive analysis utilizing multiple datasets (Math_QA, MedMCQA, and SciQ) and debate strategies (SoM-MAD and MP-MAD) highlights the complex interplay between debate formats and the performance of various LLMs. Models display unique patterns of performance changes under debate conditions, underscoring how their distinct architectures and training paradigms influence their ability to capitalize on the adversarial and iterative nature of the debates. GPT3.5-Turbo, Mixtral-8x7B-Instruct-v0.1, and Solar-10.7B-Instruct-v1.0 emerges as better models, consistently excelling across all datasets and strategies, showcasing its robustness and adaptability.

The transition from SoM-MAD to MP-MAD strategy does not significantly alter performance trends, indicating that while adding multiple personas introduces new dynamics into the debate, it does not substantially enhance the models’ performance ceilings. The analysis indicates that model effectiveness is somewhat domain-specific, with certain models achieving better results in specific datasets like SciQ and MedMCQA. Not all models derive benefit from the debate strategies, with some experiencing declines or minimal improvements. Factors such as model sensitivity to the types of errors discussed, dataset complexity, or the models’ ability to integrate adversarial feedback could contribute to these outcomes.

Additionally, the debate process itself sometimes introduces confusion among models, leading to deviations from correct answers and erratic decision-making by mediator agents, which may alter the effectiveness of these strategies.

In conclusion, while debate-based strategies have the potential to enhance LLM performance through adversarial learning and iterative refinement, their success is highly contingent upon the specific characteristics of each model and the targeted dataset. These findings underscore the necessity for customized strategies that account for the unique strengths and weaknesses of each model, aiming to optimize the benefits of adversarial learning environments. This tailored approach is essential for maximizing the efficacy of debate strategies in improving the accuracy and reliability of LLM responses.

References

- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. [MathQA: Towards interpretable math word problem solving with operation-based formalisms](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. [ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate](#). *Computing Research Repository*, arXiv:2308.07201:1–16.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. [Improving Factuality and Reasoning in Language Models through Multiagent Debate](#). *Computing Research Repository*, arXiv:2305.14325.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiwu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. [MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework](#). *Computing Research Repository*, arXiv:2308.00352.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. [Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate](#). *Computing Research Repository*, arXiv:2305.19118.
- Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023. [Training Socially Aligned Language Models on Simulated Social Interactions](#). *Computing Research Repository*, arXiv:2305.16960:1–21.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. [Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative Agents: Interactive Simulacra of Human Behavior](#). *UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, arXiv:2304.03442.
- Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. 2023. [Let Models Speak Ciphers: Multiagent Debate through Embeddings](#). *Computing Research Repository*, arXiv:2310.06272v1:1–30.
- Andries Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D. Barrett, and Arnu Pretorius. 2023. [Are we going MAD? Benchmarking Multi-Agent Debate between Language Models for Medical QA](#). *Computing Research Repository*, arXiv:2311.17371.
- Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazouze, Walter Talbott, Katherine Metcalf, Natalie Mackraz, Devon Hjelm, and Alexander Toshev. 2023. [Fine-Tuning Large Language Models with Multi-Agent Debate Supervision](#). *Computing Research Repository*, arXiv:2310.17722:1–31.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). *ArXiv*, abs/1707.06209.
- Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. [Mindstorms in Natural Language-Based Societies of Mind](#). *Computing Research Repository*, arXiv:22305.17066:1–54.

A Appendix I: Contributions of Each Team Member

This project represents a collaborative effort among all authors—R.R., V.K., and G.B.—who have contributed equally to its various phases. The development and execution of the project involved comprehensive involvement in design, proposal writing, reporting, and implementation, ensuring a robust and innovative outcome.

- **Project Design and Proposal:** The initial project design and proposal were collaboratively developed by all authors. This foundational work laid the groundwork for the subsequent phases of the project. The team collectively contributed to the drafting, revising, and finalizing of the project proposal, ensuring a strong submission.
- **Data Preprocessing:** Data preprocessing was individually handled by each author for different datasets, which were integral to the project's experimental framework. R.R. processed the MATH_QA dataset, V.K. handled the MedMCQA dataset, and G.B. took responsibility for the SciQ dataset. This division allowed for specialized focus on each dataset, ensuring optimal data quality and relevance to the project needs.
- **Implementation of LLM Code Using LangChain:** The initial implementation phase of the project utilized the LangChain interface to develop LLM code, primarily led by V.K. and G.B. Their work in this early stage was critical in setting up the basic framework upon which further functionalities were built.
- **Design of MAD Framework Using CrewAI:** R.R. took the lead in designing the MAD framework using the CrewAI platform. This involved conceptualizing and developing the framework's architecture to support the project's goals. R.R.'s initial work was pivotal for the integration of advanced features in later stages.
- **Development of MAD Framework for MP-MAD and SoM-MAD:** Subsequent to the initial designs, all authors contributed to the expansion of the MAD framework to include MP-MAD and SoM-MAD frameworks. This phase involved iterative development, testing,

and refinement, ensuring the frameworks were well-suited to the project's needs.

- **Experimentation and Testing:** The project saw extensive experimentation and testing, performed by all authors. These activities were crucial for understanding the effectiveness of different approaches and configurations. Initial testing helped identify the most promising models, which were then further refined.
- **Prompt Engineering:** One of the major challenges faced during the project was prompt engineering, where all authors played significant roles. The collective effort in designing, testing, and refining the prompts was vital for maximizing the performance of the deployed models.
- **Report Writing:** All phases of report writing—the proposal, the mid-term report, and the final report—were collaboratively written by R.R., V.K., and G.B. Each author contributed to drafting, reviewing, and finalizing the documents, ensuring thorough documentation of the project's progress and outcomes.

B Appendix II

In Appendix II, we have included the following figures:

Performance of the MP-MAD strategy in different debate rounds for various LLMs for

- Math_QA dataset: Figure 6
- MedMCQA dataset: Figure 7
- SciQ dataset: Figure 8

Furthermore, example code snippets and outputs from the Multi-Agent Debate (MAD) frameworks are provided from Figure 9 through Figure 14.

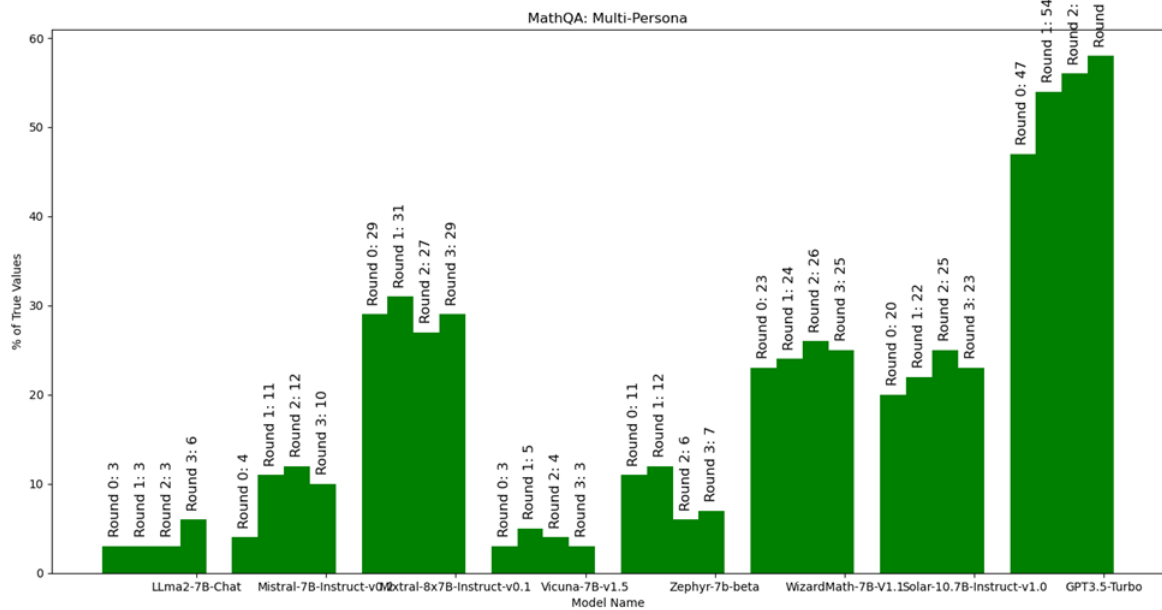


Figure 6: Performance of the MP-MAD strategy in different debate rounds for various LLMs using the Math_Q dataset, where round 0 represents the baseline single-agent performance.

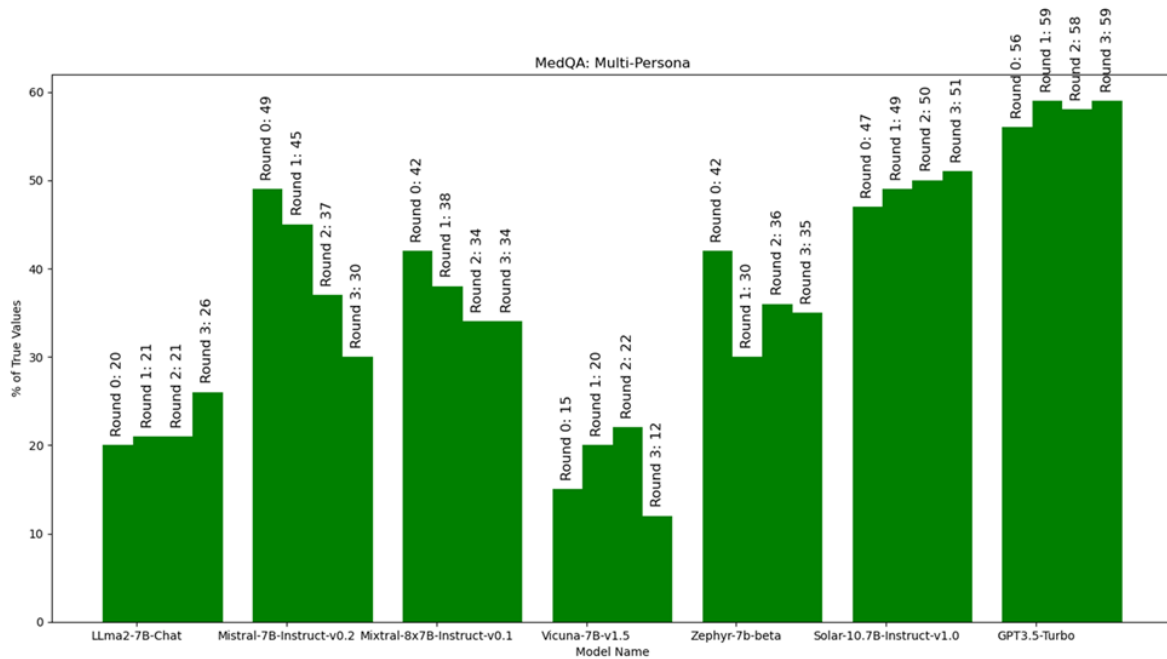


Figure 7: Performance of the MP-MAD strategy in different debate rounds for various LLMs using the MedMCQA dataset, where round 0 represents the baseline single-agent performance.

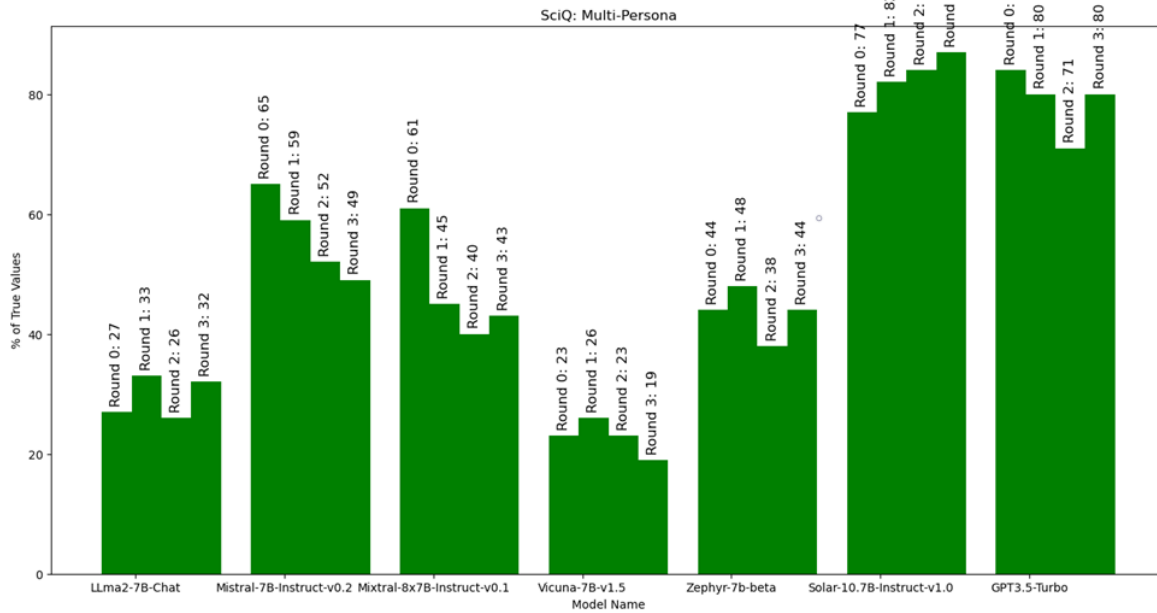


Figure 8: Performance of the MP-MAD strategy in different debate rounds for various LLMs using the SciQ dataset, where round 0 represents the baseline single-agent performance

```

> Finished chain.
[DEBUG]: == [You are an expert debator and critic] Task output: Topic: Women empowerment in western countries versus Asian countries

Context from opponent's debate: The opponent has highlighted the progress made by western countries such as Iceland, Sweden, and Norway in terms

Memory Conversions:
- Western countries have implemented policies like gender quotas, paid parental leave, and equal pay legislation to promote women empowerment.
- Asian countries have been criticized for gender discrimination, limited access to education and employment opportunities for women, and cultur

Counter Facts:
- While western countries have made progress, they still struggle with issues like the gender pay gap and underrepresentation of women in certain
- Some Asian countries have also taken steps to address gender disparities and empower women, showcasing progress in recent years.

Counter Arguments:
1. It is important to recognize that progress in women empowerment is not solely determined by policies and legislation, but also by societal at
2. While western countries may have implemented progressive policies, they still face challenges in achieving true gender equality and empowerin

Summary:
In conclusion, the debate on women empowerment in western countries versus Asian countries is complex and multifaceted. While western countries

```

Figure 9: MAD output for a representative question: "Women empowerment in Western vs Asian countries"

```

def task_debate_student1(science_question, options, agent, debate_student2):
    return Task(
        description= dedent(f"""
        Give the correct answer label from the options given: {options} for the medical science question: {science_question}.
        Give detailed reasonings and arguments for selecting your answer.
        If your friend is debating with his arguments and reasonings, you must debate with your arguments and reasonings for your answer.
        His debate arguments are given as context.
        If he is wrong, correct him and give correct answer.
        If your friend's debate provide you new insights, update your reasoning about answering the medical science question.
        You may also use your previous conversations from memory or available contexts for your argumnets and reasonings for your answer."""),
        expected_output= dedent(f"""
        Detailed summary of your reasonings and arguments for selecting the answer for the given medical science question.
        If you have any debate with other student, give your debate arguments and reasonings.
        Do not forget to include your answer and the correct answer label from the options given {options}.
        Example:
        Correct answer label:
        Reason/argument: ""
        """),
        context = [debate_student2],
        agent=agent,
        output_file="debate_student1.txt",
    )

```

Figure 10: Prompt for "Student2" in the CrewAI Task Setup Utilizing the SoM-MAD Strategy for Science Question Answering"

Student1

Natural selection is the term for the process in which living things with beneficial traits produce more offspring than others do. This process is driven by the environment and the struggle for survival, and results in the adaptation of species to their environment. Natural survival, natural process, and natural variety do not accurately describe this specific process. Therefore, the correct answer label for the given science question is "natural selection."

Student2

After reviewing my friend's arguments and reasonings, I understand that he believes that natural survival is the term for the process in which living things with beneficial traits produce more offspring than others do. He argues that this process is driven by the environment and the struggle for survival, and results in the adaptation of species to their environment. However, I counter that natural selection is the term for the process in which living things with beneficial traits produce more offspring than others do. This process is driven by the environment and the struggle for survival, and results in the adaptation of species to their environment. Natural survival, natural process, and natural variety do not accurately describe this specific process. Therefore, the correct answer label for the given science question is "natural selection."

Figure 11: Example output of the SoM-MAD strategy for a debate between two "Student" agents in solving a science question

```
def task_grading(math_question, agent, task_context_solving):
    return Task(
        description=dedent(f"""
            For the solution provided for the math problem: {math_question}, debate on the student's answer. Think step by step
            Your response must start with "I disagree with you. here is my reasons/arguments" as you are a great critic of students
            Use previous conversations and contexts available for debating.
            You must review the solution, give feedback to student and provide correct solutions.
            If the solution is wrong, give correct solution with explanation.
            Provide alternative pathways or easy methods to solve the math question if any.
            If you found the answer is wrong, ask the student to improve or correct his answer using your comments/feedback
            including alternative pathways as suggested.""",
        ),
        expected_output = dedent(f""" \
            your response must start with "I disagree with you. here is my reasons/arguments:"
            A detailed feedback of student's solution. Give alternative pathways to solutions and provide correct solutions.""",
        ),
        context = [task_context_solving],
        agent=agent,
        output_file = 'grading_task.txt'
    )
```

Figure 12: Prompt for the "Professor" in the CrewAI Task Setup Employing the MP-MAD Strategy for Math Reasoning"

```
def task_debate_student2(science_question, options, agent, debate_student1):
    return Task(
        description=dedent(f"""
            You are big critic of your friend and always debate with him with your counter arguments and reasonings on the given medical science question: {science_question}
            You always disagree with your friend and provide alternative pathways to answer the question.
            Your response must start with "I disagree with you. Here are my reasons/arguments"
            After debate with your friends, give the correct answer and select the correct answer label from the options {options}
            You may also use your previous conversations and contexts from memory for your arguments and reasonings for your answer.
            """,
        ),
        expected_output = dedent(f""" \
            A detailed summary of your reasonings and arguments for selecting the answer for the given medical science question.
            Give your debate arguments and reasonings that supports your debate.
            Do not forget to include your answer and the correct answer label from the given options list {options}.
            Example:
            Correct answer label:
            Reason/argument: """,
        ),
        context = [debate_student1],
        agent=agent,
        output_file = 'debate_student2.txt'
    )
```

Figure 13: Prompt for the "Professor" in the CrewAI Task Setup Employing the SoM-MAD Strategy for Science Reasoning

Student

After considering the feedback from my professor, I revised my approach to solving the math problem.

- Let the distance from point a to point b be d km.
- The car travels the first half of the trip at a speed of 80 km/hr, which means it covers d/2 distance at that speed.
- The time taken to cover the first half of the trip is (d/2) / 80 = d/160 hours.
- The remaining distance to cover in the second half of the trip is also d/2 km.
- We know that the average speed for the whole trip is 60 km/hr.
- Using the formula for average speed = total distance / total time, we have: $60 = d / ((d/2) / 80 + (d/2) / x)$, where x is the speed in the second half of the trip.

After solving the equation, we find that $x = 48$ km/hr.

Therefore, the correct speed of the car in the second half of the trip is 48 km/hr. Taking into account the different speeds and distances covered in each half of the trip is crucial when calculating the average speed.

I have incorporated the correct calculations and explanations provided by my professor to arrive at the final answer of 48 km/hr for the speed of the car in the second half of the trip. Thank you for the valuable feedback and guidance.

Professor:

I disagree with you. Here is my reasons/arguments:

The student's approach to the problem is partially correct, but there are some errors in the calculations. Let's correct and provide feedback on the solution:

- The student correctly identified the distance from point a to point b as d km.
- The car travels the first half of the trip at a speed of 80 km/hr, covering d/2 distance at that speed.
- The time taken to cover the first half of the trip is (d/2) / 80 = d/160 hours.
- The remaining distance to cover in the second half of the trip is also d/2 km.
- The average speed for the whole trip is 60 km/hr.
- The formula for average speed = total distance / total time is correct, but the setup of the equation should be: $60 = d / ((d/2) / 80 + (d/2) / x)$, where x is the speed in the second half of the trip.

After solving the corrected equation, we find that $x = 96$ km/hr, not 48 km/hr as the student calculated.

Therefore, the correct speed of the car in the second half of the trip is 96 km/hr, not 48 km/hr. It is important to carefully consider the different speeds and distances covered in each half of the trip when calculating the average speed.

I have provided the necessary corrections and feedback to the student's solution. It is crucial to pay attention to the details and ensure accurate calculations to arrive at the correct answer. Thank you for your attention to these points.

Figure 14: Example output of the MP-MAD strategy for a debate between 'Student' and 'Professor' agents in solving a math question