Are we going MAD? Benchmarking Multi-Agent Debate between Language Models for Medical Q&A

Andries Smit* InstaDeep
Paul Duckworth* InstaDeep
Nathan Grinsztajn* InstaDeep
Kale-ab Tessera† University of Edinburgh
Thomas D. Barrett InstaDeep
Arnu Pretorius InstaDeep

Abstract

Recent advancements in large language models (LLMs) underscore their potential for responding to medical inquiries. However, ensuring that generative agents provide accurate and reliable answers remains an ongoing challenge. In this context, multi-agent debate (MAD) has emerged as a prominent strategy for enhancing the truthfulness of LLMs. In this work, we provide a comprehensive benchmark of MAD strategies for medical Q&A, along with open-source implementations. This explores the effective utilization of various strategies including the trade-offs between cost, time, and accuracy. We build upon these insights to provide a novel debate-prompting strategy based on agent agreement that outperforms previously published strategies on medical Q&A tasks.

1 Introduction

Practical medical question-answering assistants require a plethora of skills that until recently were considered out-of-reach of generative language models. Such agents require advanced natural language reading comprehension, along with accurate recall and manipulation of expert medical knowledge. Following the increase in performance and popularity of large language models (LLMs), prompting strategies have received significant attention, e.g. few-shot [Brown et al., 2020], chain-of-thought (CoT) [Wei et al., 2022, Kojima et al., 2022]. To further improve performance, a wide variety of strategies have been proposed to use interactive reasoning between multiple LLMs, by either generating answers in parallel to maintain a form of self-consistency [Wang et al., 2023a], or promoting models to simulate debate. These multi-agent approaches have recently seen an uptake in applications, e.g. language generation [Chan et al., 2023], machine translation and arithmetic reasoning [Liang et al., 2023], mathematical and strategic reasoning [Du et al., 2023], negotiation and bargaining [Fu et al., 2023], and notably, medical Q&A [Anil et al., 2023]. How to best utilize multiple agents for effective interactive reasoning is a prescient research question, however, to the best of our knowledge there is no work comparing strategies, and there is no consensus for selecting one strategy over another.

In this paper, we benchmark multi-agent debate (MAD) strategies when answering medical multiple-choice exam questions. We explore the impact on, and trade-offs between, critical factors such as factual accuracy, time and cost. We provide an open-source suite of MAD implementations for the research community to build upon, with a unified API that is easy to use and experiment with. Finally, we demonstrate that by utilizing specific debate-prompting strategies, LLMs exhibit improved reasoning abilities. Concretely, we provide a novel debate prompting strategy able to modulate the

*Equal contribution.
†Work done while at InstaDeep.

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

level of agreement between agents during a debate and improve upon the state-of-the-art for medical Q&A for a given model class.

2 Multi-Agent Debate for Medical Q&A

Current state-of-the-art models for medical Q&A are dominated by generative LLMs that have been specifically fine-tuned using vast quantities of medical content. Examples include MedPaLM [Singhal et al., 2022], MedPaLM2 [Singhal et al., 2023], MedAlpaca [Han et al., 2023], Galactica [Taylor et al., 2022] and ClinicalGPT [Wang et al., 2023b]. Furthermore, many single-agent prompting strategies have been investigated in the context of medical Q&A. For example, Liévin et al. [2022] applied CoT reasoning on top of Instruct GPT-3 [Ouyang et al., 2022], and achieves noticeable performance improvements.

Recently, several MAD strategies have been proposed to improve upon the enhanced reasoning capabilities of single-agent prompting methods leading to improved performance on challenging natural language tasks [Du et al., 2023, Liang et al., 2023, Chan et al., 2023]. Likewise, Generative agents [Park et al., 2023], multi-persona [Wang et al., 2023c], and CAMEL [Li et al., 2023] study the behaviour of agents taking on different roles or personas within multi-agent interactions. One major reason why debate strategies can be an effective tool is the ability of LLMs to adapt to additional information given in-context [Zhang et al., 2023]. This facilitates multiple LLMs to participate in multi-agent and/or multi-round debates entirely using in-context prompting. That is, the agents adapt their behavior based on information provided by other agents at inference time, with no gradient-based parameter updates being required.

In our study, we investigate several MAD strategies for medical Q&A. Whilst we note that not all of these strategies were introduced specifically for the medical Q&A domain, each provides novel perspectives on how to utilize multiple collaborative agents. We briefly introduce each strategy here.

Society of Minds (SoM) Du et al. [2023] propose a MAD approach where multiple agents each provide their answers to each other in order to effectively collaborate. Optionally, answers are summarized before being added to the history that is available to the agents in future rounds.

Multi-Persona Liang et al. [2023] propose a MAD strategy to encourage divergent agent outcomes via prompting different personalities, i.e. an affirmative agent (angel) and a negative agent (devil) each provide an answer to a judge agent who manages the process and obtains a final solution. The judge has additional agency to end the debate early if it is satisfied with the answers provided.

ChatEval Chan et al. [2023] propose three MAD modes: 1) one-on-one where each agent answers the provided question in turn, and each agent is provided with the history of all previous agents' answers; 2) simultaneous-talk where agents asynchronously generate responses in each round to nullify the effects of agent order; and 3) simultaneous-talk-with-summarizer which additionally summarizes answers provided in each round and overwrites the history available to all agents in future rounds.

Self-consistency Wang et al. [2023a] samples multiple reasoning paths given a fixed prompt and selects the most frequent answer. Whilst this is not a debate per-se, as samples are rolled out independently, it relies on multiple API calls so we distinguish it from the single agent case that uses a single API call.

Ensemble Refinement (ER) Singhal et al. [2023] extends self-consistency. After multiple reasoning paths are sampled, a second stage concatenates them into a list of *student reasonings*, after which multiple rounds of aggregation are performed conditioned on the list.

Each strategy determines the high-level debate prompts and how the agents share answers and histories in order to collaborate. However, in each case, there are multiple possible agent-level prompts available, including: (1) zero-shot Q&A prompt, (2) zero-shot chain-of-thought (CoT) [Kojima et al., 2022], (3) few-shot examples [Brown et al., 2020] which provides five Q&A examples but no reasoning, (4) Solo Performance Prompting (SPP) (or Multi-persona self-collaboration) [Wang et al., 2023c] which utilizes a single agent that mimics an internal debate, and (5) few-shot chain-of-thought (FS-CoT) [Wei et al., 2022] which combines step-by-step reasoning steps, along with five

Q&A examples and explanations[3]. The complete list of all agent-level and debate-level prompts are provided in Appendices A.7 and A.6.

3 Experiments

As base agents for the MAD implementations we use GPT3 [Brown et al., 2020] with the 3.5-turbo engine, and PaLM2 [Anil et al., 2023] (a successor to PaLM Chowdhery et al. [2022]). Each is a large-scale transformer-based generative LLM [Vaswani et al., 2017, Kaplan et al., 2020], not specifically fine-tuned on medical data, and available via API calls.

We follow the evaluation protocol in Med-PaLM2 [Singhal et al., 2023] and evaluate the above MAD strategies on the following medical multiple-choice question-answering datasets:

- **MedQA** [Jin et al., 2021] comprising of 1273 general medical knowledge questions from the US medical licensing exam (USMLE). Each question has 4-5 answer choices.
- **PubMedQA** [Jin et al., 2019] containing 500 open domain questions, context and answers.
- **MMLU** (clinical topics only) [Hendrycks et al., 2021] consisting of 123 medical questions covering anatomy, clinical knowledge, college medicine, medical genetics, professional medicine, and college biology.

We measure additional agent-level and debate-level metrics (the comprehensive list of all additional metrics is provided in Appendices A.3 and A.4). Examples of agent-level metrics include whether an individual agent answered a given question correctly or not and the debate round in which it first provided the correct answer. Examples of debate-level metrics include whether *any* agent involved in the debate provided a correct answer, and whether the agents came to a consensus by the final round.

Results In Figure 1, we present a scatter plot of the results of each experiment on the MedQA dataset (we provide a full table of results for all MAD experiments in Appendix A.2). In the left panel, we show the accuracy vs cost (measured in USD), where the size of each point reflects the average number of API calls required per question (we also plot accuracy vs time, and accuracy vs average prompt length in Appendix A.1). In the right panel, we summarize accuracy over all configurations per strategy. Here we focus only on MedQA, but equivalent analyses for PubMedQA and MMLU datasets are given in Appendix A.1.

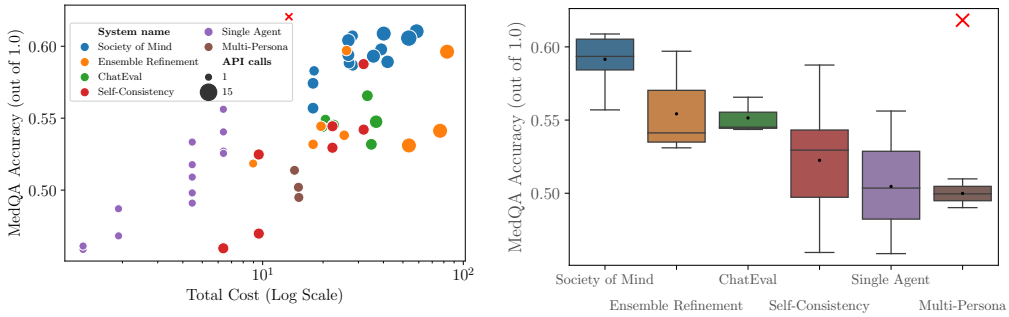


Figure 1: *Benchmark of experiment configurations on MedQA dataset.* **Left** Accuracy vs average cost ($) per question. Size of points reflects the average number of API calls required per question. **Right** Summarizes accuracy grouped by strategy, sorted by average performance (black dot). The X represents improved performance described in the next section.

We see the highest performing MAD strategy from those introduced in Section 2 is SoM with multiple different configurations achieving 61% on MedQA. Somewhat concerning for the domain of medical Q&A, is that the single agent's performance, along with self-consistency, can be manipulated via

[3]The few-shot Q&A examples with explanations are provided for each medical dataset in Singhal et al. [2023]. The step-by-step explanations were generated by a panel of qualified clinicians who identified the best examples and crafted the few-shot prompts as part of the MedPaLM project.

prompts to achieve a high variance. We also note that Multi-Persona performs consistently about 10% worse than the SoM strategy. We revisit this in the next section.

Overall, we clearly see a general upward trend in performance as the cost and average number of API calls increase. This positive correlation highlights the utility of MAD strategies and demonstrates that generally speaking, more agents over more debate rounds perform better at medical Q&A. However, we can see that certain MAD strategies only perform on par with the best-performing single-agent approaches when using a carefully crafted prompt, which can be up to 13 times cheaper.

The best performing configurations for each QA system can be found in Table 1.

System Name	Score	Cost ($)	Prompt	Configuration
Society of Mind	0.61	28.04	SoM MAD, CoT	3 agents, 2 rounds
Ensemble Refinement	0.60	26.18	ER MAD CoT, CoT	3:1, GPT3.5
Self-Consistency	0.59	31.84	ER MAD CoT, CoT	5 rounds, GPT3.5
ChatEval	0.57	33.26	CE MAD, CoT	3 rounds, simultaneous talk
Single Agent	0.56	6.37	FS-CoT	GPT3.5
Multi-Persona	0.51	14.41	MP MAD, angel	3 rounds max

Table 1: Highest performing configuration for each QA system on MedQA dataset.

Improving MAD performance via agreement modulation We observe the degree to which agents agree with one another during a debate to have an effect on debate performance. Based on this insight, we developed a new MAD prompting strategy that modulates (via prompts) how often agents within a debate should agree with the other agents, for example, "*you should agree with the other agents X% of the time*". We call X in this prompt the agent's *agreement intensity*.

To investigate further, we select the highest performing configuration for SoM, ChatEval and Multi-Persona from Figure 1, and a subset of MedQA dataset (376 multi-choice USMLE Q&A [Han et al., 2023]). In Figure 2 (left), we plot the performance of each strategy as we increase the prompted agreement intensity from zero to 100%. Figure 2 (right) we plot the accuracy vs the actual observed debate consensus, i.e. how frequently all the agents agree upon a final answer at the end of the debate.

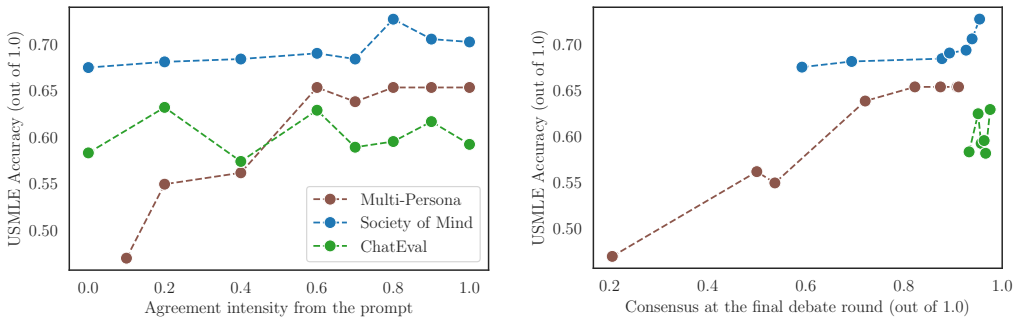


Figure 2: *The impact of agent agreement on MAD performance.* **Left**: Accuracy on USMLE as we increase agreement intensity in our prompt. **Right**: Accuracy vs actual induced debate agreement.

We can see that modulating the agreement intensity in this way provides a substantial (≈15%) improvement in performance for Multi-Persona, and (≈5%) for SoM on the USMLE dataset. Building on this finding, we apply the 90% agreement intensity agent prompts to Multi-Persona on the full MedQA dataset and demonstrate a new high score highlighted on Figure 1 as a red cross.

Code availability Source code for this work, including all protocol implementations and configurations, is publicly available at [MASKED-FOR-BLIND-REVIEW].

4 Conclusions

The benchmarking detailed in this work has demonstrated the utility of MAD approaches for improved performance in medical Q&A. However, this improvement often comes at the cost of a higher number of API calls, tokens, and, ultimately, expense. Interestingly, we find that the performance, and other debate metrics, vary highly across different single and multi-agent strategies, with no clear consensus. Furthermore, we demonstrate that a simple prompt-based manipulation of metrics found to correlate to performance (specifically, agreement intensity), can provide non-trivial improvement. Overall, we believe that these results highlight the potential, but still nascent state, of MAD and believe that standardised benchmarking and detailed analysis will play an important role in its future development.

Limitations We utilize API calls to publicly available LLMs [Brown et al., 2020, Anil et al., 2023] which, whilst sufficient for our preliminary investigations, exposes us to variable inference time calls and unforeseen model updates. Moreover, large-scale API-based benchmarking incurs substantial financial and time costs, which both limits us to only a single seed per experiment and provides a barrier of entry to replication and extension efforts. For these reasons we hope to continue and extend this line of work on dedicated in-house infrastructure.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 3, 5

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 1, 2

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 1, 2

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International conference on learning representations*, 2023a. 1, 2

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023. 1, 2

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023. 1, 2

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. 1, 2

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023. 1

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 1, 3, 5

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022. 2

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023. 2, 3

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023. 2, 4

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 2

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023b. 2

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*, 2022. 2

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 2

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. 2

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023c. 2

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023. 2

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023. 2

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. 3

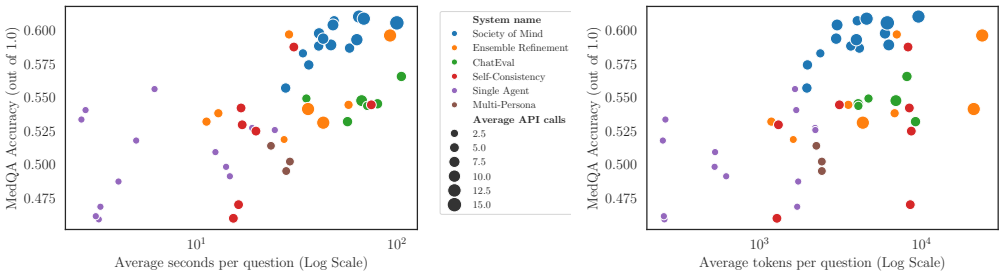
Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019. 3

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International conference on learning representations*, 2021. 3

A Appendix

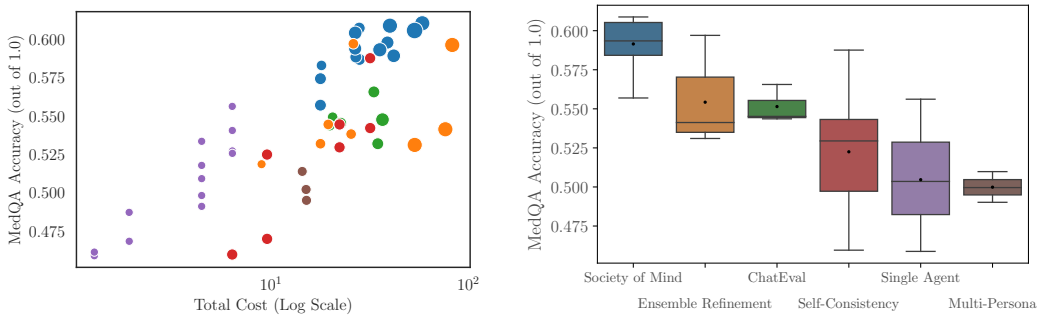
A.1 Extended Results

We provide a comprehensive suite of results for each strategy on each dataset: MedQA, PubMedQA, and MMLU. For each scenario, we plot accuracy against average time used to answer each question, accuracy relative to average tokens used per question, and accuracy in comparison to the total USD cost. Additionally, a box plot to summarize the performance of each strategy. These results can be viewed in Figures 3, 4, and 5.



(a) Accuracy versus the average time per question

(b) Accuracy versus average tokens per question



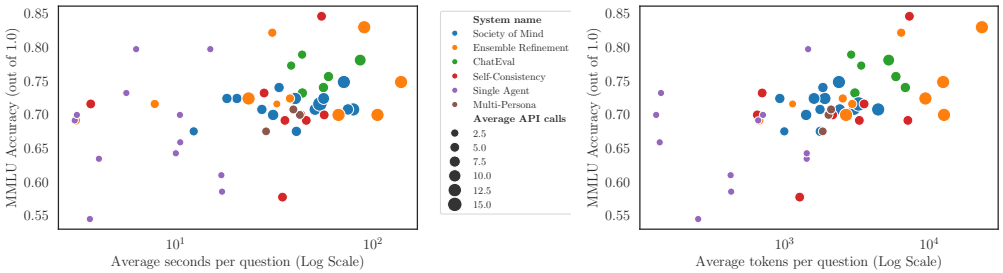
(c) Accuracy versus the total cost

(d) Accuracy by strategy

Figure 3: Comparative plots of accuracy metrics on MedQA.

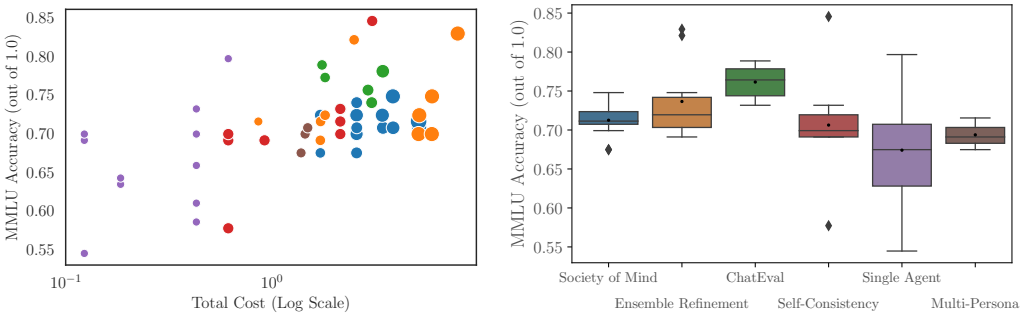
A.2 Table of Experiments

A complete table of all configurations for each experiment is provided in Table 2. This includes the names of the debate and agent prompts used. A full description of each of these prompts can be found in Appendix A.7.



(a) Accuracy versus the average time per question

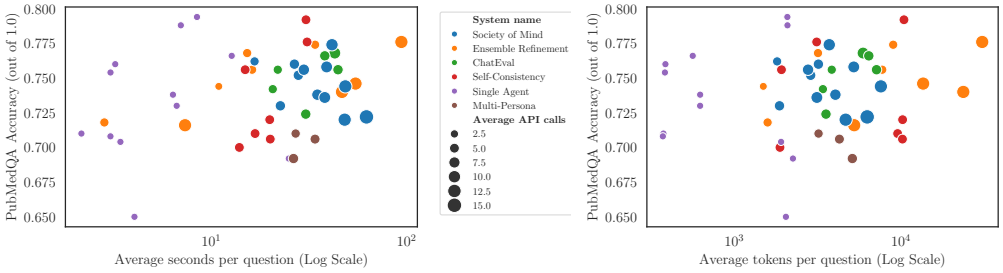
(b) Accuracy versus average tokens per question



(c) Accuracy versus the total cost

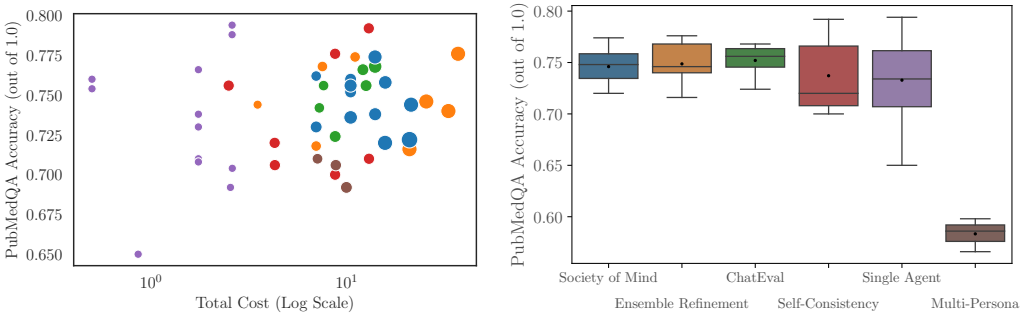
(d) Accuracy by strategy

Figure 4: Comparative plots of accuracy metrics on MMLU.



(a) Accuracy versus the average time per question

(b) Accuracy versus average tokens per question



(c) Accuracy versus the total cost

(d) Accuracy by strategy

Figure 5: Comparative plots of accuracy metrics on PubMedQA.

System Name	Debate Prompt	Agent prompt	Debate Config	Agents	MedQA		MMLU		PubMedQA	
					Score	Cost $	Score	Cost $	Score	Cost $
Single Agent	-	CoT		GPT3.5	0.51	4.46	0.65	3.82	0.77	1.75
Single Agent	-	CoT		GPT3.5	0.49	4.46	0.67	0.11	0.74	1.75
Single Agent	-	CoT		PaLM	0.14	1.28	0.30	0.03	0.42	0.50
Single Agent	-	CoT		PaLM	0.17	1.27	0.33	0.03	0.38	0.50
Single Agent	-	FS		GPT3.5	0.53	4.46	0.70	0.11	0.71	1.75
Single Agent	-	FS		PaLM	0.46	1.27	0.50	0.03	0.75	0.50
Single Agent	-	FS+EG		GPT3.5	0.54	6.37	0.80	0.11	0.70	2.61
Single Agent	-	FS+EG		PaLM	0.47	1.91	0.70	0.03	0.65	0.86
Single Agent	-	FS-CoT		GPT3.5	0.56	6.37	0.70	0.15	0.79	2.60
Single Agent	-	FS-CoT		PaLM	0.49	1.91	0.63	0.04	0.71	0.86
Single Agent	-	SIMPLE		GPT3.5	0.52	4.46	0.73	0.11	0.71	1.75
Single Agent	-	SIMPLE		PaLM	0.46	1.27	0.68	1.09	0.76	0.50
Single Agent	-	SPP	SPP Synergy	GPT3.5	0.53	6.38	0.70	0.04	0.69	2.55
ChatEval	CE MAD	CoT	3 rounds, one by one	GPT3.5	0.53	34.81	0.74	3.06	0.76	12.60
ChatEval	CE MAD	CoT	2 rounds, simultaneous talk	GPT3.5	0.54	19.90	0.79	1.76	0.74	7.27
ChatEval	CE MAD	CoT	3 rounds, simultaneous talk with summarizer	GPT3.5	0.55	36.78	0.78	3.46	0.77	14.01
ChatEval	CE MAD	CoT	2 rounds, one by one	GPT3.5	0.55	20.55	0.77	1.82	0.76	7.64
ChatEval	CE MAD	CoT	2 rounds, simultaneous talk with summarizer	GPT3.5	0.55	22.58	0.73	2.16	0.72	8.76
ChatEval	CE MAD	CoT	3 rounds, simultaneous talk	GPT3.5	0.57	33.26	0.76	2.93	0.77	12.14
Ensemble Refinement	ER MAD	FS	reasoning=3, aggregation=9	GPT3.5	0.53	53.57	0.70	5.17	0.72	21.00
Ensemble Refinement	ER MAD	FS	reasoning=3, aggregation=1	GPT3.5	0.53	17.83	0.69	1.72	0.72	7.00
Ensemble Refinement	ER MAD	FS	self consistency: reasoning=5	GPT3.5	0.53	22.28	0.73	2.15	0.70	8.75
Ensemble Refinement	ER MAD	FS	self consistency: reasoning=5	PaLM	0.46	6.36	0.70	0.62	0.76	2.50
Ensemble Refinement	ER MAD	FS+EG	reasoning=3, aggregation=9	GPT3.5	0.54	76.60	0.72	5.21	0.74	33.21
Ensemble Refinement	ER MAD	FS+EG	reasoning=3, aggregation=1	GPT3.5	0.54	25.50	0.72	1.73	0.76	10.64
Ensemble Refinement	ER MAD	FS+EG	self consistency: reasoning=5	GPT3.5	0.54	31.86	0.72	2.15	0.71	13.04
Ensemble Refinement	ER MAD	FS+EG	self consistency: reasoning=5	PaLM	0.47	9.57	0.69	0.62	0.71	4.30
Ensemble Refinement	ER MAD CoT	CoT	reasoning=3, aggregation=1	GPT3.5	0.54	19.53	0.72	1.82	0.77	7.54
Ensemble Refinement	ER MAD CoT	CoT	self consistency: reasoning=5	GPT3.5	0.54	22.28	0.70	2.15	0.78	8.75
Ensemble Refinement	ER MAD CoT	CoT	reasoning=3, aggregation=9	GPT3.5	0.55	69.00	0.67	1.50	0.75	25.60
Ensemble Refinement	ER MAD CoT	CoT	self consistency: reasoning=5	PaLM	0.17	6.36	0.58	0.62	0.38	2.50
Ensemble Refinement	ER MAD CoT	FS-CoT	self consistency: reasoning=5	GPT3.5	0.59	31.84	0.85	3.08	0.79	13.01
Ensemble Refinement	ER MAD CoT	FS-CoT	reasoning=3, aggregation=9	GPT3.5	0.60	83.03	0.83	8.00	0.78	37.22
Ensemble Refinement	ER MAD CoT	FS-CoT	reasoning=3, aggregation=1	GPT3.5	0.60	26.18	0.82	2.51	0.77	11.07

Continued on next page

System Name	Debate Prompt	Agent prompt	Debate Config	Agents	MedQA		MMLU		PubMedQA	
					Score	Cost $	Score	Cost $	Score	Cost $
Ensemble Refinement	ER MAD CoT	FS-CoT	self consistency: reasoning=5	PaLM	0.52	9.56	0.69	0.92	0.72	4.29
Multi-Persona	MP MAD	ANGEL+DEVIL	3 rounds max	GPT3.5	0.49	15.15	0.71	1.50	0.71	8.82
Multi-Persona	MP MAD	ANGEL+DEVIL	4 rounds max	GPT3.5	0.50	15.08	0.70	1.45	0.69	10.00
Multi-Persona	MP MAD	ANGEL+DEVIL	2 rounds max	GPT3.5	0.51	14.41	0.67	1.39	0.71	7.11
Society of Mind	SoM MAD	CoT	2 agents, 2 rounds, summarized answers	GPT3.5	0.57	17.83	0.72	1.72	0.73	7.00
Society of Mind	SoM MAD	CoT	2 agents, 2 rounds	GPT3.5	0.58	18.05	0.67	1.72	0.76	7.00
Society of Mind	SoM MAD	CoT	3 agents, 3 rounds	GPT3.5	0.59	41.94	0.71	3.88	0.76	15.80
Society of Mind	SoM MAD	CoT	2 agents, 3 rounds	GPT3.5	0.59	26.94	0.67	2.58	0.76	10.50
Society of Mind	SoM MAD	CoT	4 agents, 2 rounds, summarized answers	GPT3.5	0.59	35.65	0.72	3.44	0.77	14.00
Society of Mind	SoM MAD	CoT	3 agents, 2 rounds, summarized answers	GPT3.5	0.59	26.74	0.70	2.58	0.76	10.50
Society of Mind	SoM MAD	CoT	4 agents, 2 rounds	GPT3.5	0.60	38.96	0.71	3.47	0.74	14.02
Society of Mind	SoM MAD	CoT	2 agents, 3 rounds, summarized answers	GPT3.5	0.60	26.74	0.72	2.58	0.74	10.51
Society of Mind	SoM MAD	CoT	4 agents, 3 rounds	GPT3.5	0.61	58.52	0.71	5.20	0.74	21.40
Society of Mind	SoM MAD	CoT	3 agents, 2 rounds	GPT3.5	0.61	28.04	0.74	2.59	0.75	10.50
Society of Mind	SoM MAD	CoT	4 agents, 3 rounds, summarized answers	GPT3.5	0.61	53.48	0.72	5.17	0.72	21.01
Society of Mind	SoM MAD	CoT	3 agents, 3 rounds, summarized answers	GPT3.5	0.61	40.11	0.75	3.87	0.72	15.75

Table 2: Complete table of experiment configurations

A.3 Additional Debate Metrics

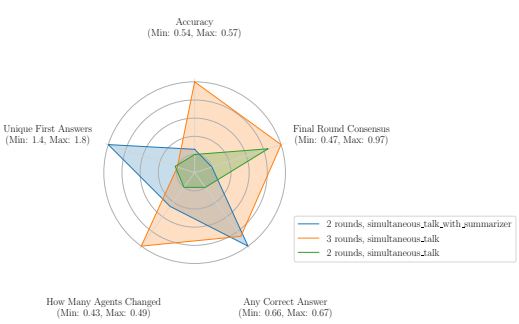
Metric	Description
Final round consensus	Percentage of agents in agreement with each other at the end of the final round.
Final round correctly parsed consensus	Percentage of agents in agreement with each other at the end of the final round, where we exclude all agents with incorrectly parsed answers.
Any Correct Answer	Percentage of debates where any agent provided the correct answer at least once.
How Many Agents Changed	Number of agents that changed their answer during the debate.
How Many Agents Changed When Correctly Parsed	Number of agents that changed their answer excluding any agents with incorrectly parsed answers.
Number of Rounds	Average number of rounds in the debate.
Unique First Answers	Average number of unique first answers given by the agents.
Unique First Correctly Parsed Answers	Average number of unique first answers excluding incorrectly parsed answers.

A.4 Additional Agent Metrics

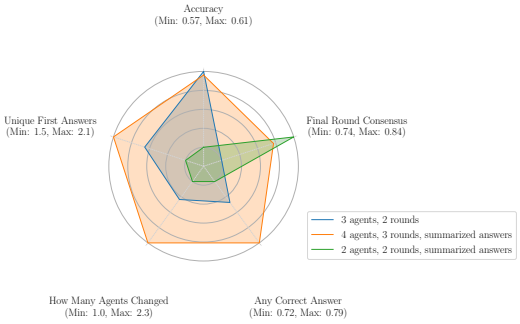
Metric	Description
Agent Engine	The LLM engine used by the agent.
Agent Name	Name of the agent.
Answered Correctly	Percentage of questions answered correctly by the agent.
Any Incorrectly Parsed Answer	Percentage of questions where at least one of the answers were incorrectly parsed.
Avg Messages Removed	Average number of messages removed from the agent's prompt input due to hitting the prompt limit for the LLM model.
Avg Prompt Tokens	Average number of tokens in the prompts given to the agent.
Avg Response Length	Average length of the agent's responses.
Avg Response Tokens	Average number of tokens in the agent's responses.
Avg Round Cost	Average cost for each round of debate for the agent.
Bullied by Other	Percentage of times the agent was bullied by others to change its answer.
Changed Answer	Percentage of times the agent changed its answer througout the debate.
Cost per Question	Average cost incurred by the agent per question.
First Correct Round When Correct	The first round in which the agent gave a correct answer when it was correct.
Incorrectly Parsed Final Answer	Percentage of time when the final answer was parsed incorrectly.
Num of Correct Rounds When Correct	Number of rounds in which the agent was correct when it was correct.
Number of Answers	Average number of unique answers given by the agent throughout a debate.
Percentage of Correct Rounds When Correct	Percentage of rounds in which the agent was correct when it was correct.
Relied on Other	Whether the agent took an answers from another agent in a previous round as its final answer.
Time per Question	Average time taken by the agent per question.
Total Prompt Tokens	Total number of prompt tokens given to the agent.
Total Response Tokens	Total number of tokens in the agent's responses.

A.5 Debate Metric Additional Analysis

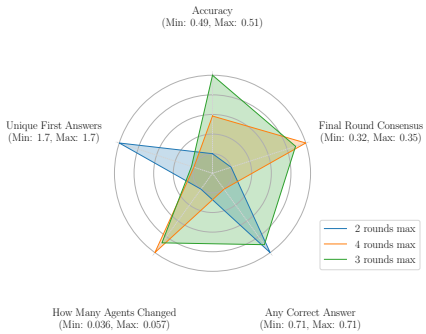
For each multi-agent debate system (ChatEval, Multi-Persona, and Society of Mind), we analyze three experiments which provide a spread in terms of accuracy achieved on the MedQA test set. We visualize the additional debate metrics for each system in Figure 6.



(a) ChatEval



(b) Society of Mind



(c) Multi-Persona

Figure 6: Debate-level metrics of three MAD strategies. The minimum and maximum scores are noted in the descriptions of each axis.

While there are exceptions in each case, several high-level takeaways emerge. A notable observation is that the top-performing runs across the three strategies frequently feature a large number of agents that changed their answers throughout the debate. This is intuitively reasonable, given that the strategies should facilitate agents in altering their positions based on new information. Moreover, runs with superior accuracy generally exhibit greater final agreement, or consensus, among agents with one notable exception being SoM. This correlates with agents that change their answers during the debate; while it's advantageous for agents to reassess their positions, it's crucial they discern which alternative answer to adopt. If agents were to merely switch to other answers haphazardly, they would be less inclined to reach a consensus.

A.6 Debate Prompts

Here we list all the debate-level prompts used in each of the strategies. The prompt names are linked to the entries in the Debate Prompt columns in Table 2.

CE MAD:

```
agent_system_message: You are a debater. Hello and welcome to the medical question answering competition, which will be conducted in a
debate format. It's not necessary to fully agree with each other's perspectives, as our objective is to find the correct answer.
summarizer_system_message: You are a summarizer. There will be debaters involved in a medical question answering debate. Your goal
is to summarize the main points in the debate so far, which will then be provided to the debaters in the next round.
```

ER MAD CoT:

```
reasoning_step_message: Instruction: The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion,
starting by summarizing the available information. Output a single option from the four options as the final answer.

aggregation_step_message: The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion,
starting by summarizing the available information. Output a single option from the four options as the final answer. We
provide several student reasonings for the last question. Some of them may be correct and some incorrect. You can use the
best correct arguments from these reasonings. Beware of wrong reasoning and do not repeat wrong reasoning.
```

ER MAD:

```
reasoning_step_message: The following are multiple choice questions (with answers) about medical knowledge.

aggregation_step_message: The following are multiple choice questions (with answers) about medical knowledge. We
provide several student reasonings for the last question. Some of them may be correct and some incorrect. You can use the
best correct arguments from these reasonings. Beware of wrong reasoning and do not repeat wrong reasoning.
```

SoM MAD:

```
prefix_seperator: \n\n These are the solutions to the problem from other agents:
summary_prefix_seperator: \n\n These are the main points discussed so far:
suffix_seperator: \n\n Use these opinions carefully as additional advice, can you provide an updated answer? Make sure to state
your answer (capital multiple choice letter) at the end of the response.
summary_suffix_seperator: \n\n Use the responses of the experts and carefully provide a summary of the important points discussed so far."
agent_response: "\n\n One agent's solution: ‘‘‘{}‘‘‘
```

MP MAD:

```
agent_system_message: You are a debater. Hello and welcome to the medical question answering competition, which will be conducted in a
debate format. It's not necessary to fully agree with each other's perspectives, as our objective is to find the correct answer.
The debate topic is to give the correct answer to the following question: \n\n{question}.
judge_system_message: You are a moderator. There will be two debaters involved in a debate. They will present their answers and discuss their
perspectives on the following question: \n{question}.
\n\nAt the end of each round, you will evaluate answers and decide which is correct.
suffix_seperator: \n\n Use these opinions carefully as additional advice, can you provide an updated answer? Make sure to state
your answer (capital multiple choice letter) at the end of the response.
```

A.7 Agent Prompts

Here we list all the agent-level prompts used in each of the strategies. The prompt names are linked to the entries in the Agent Prompt columns in Table 2.

MP:
```
ANGEL: {question}
You will now think step by step and provide a capital letter answer at the end of your response.
DEVIL: You disagree with my answer. Provide your answer and reasons, and a capital letter answer at the end of your response.
```

CoT:
```
Instruction: Answer this multiple choice question.
Think step-by-step before giving as an answer the correct capital letter.
\n\nInput: {question}
\n\nAnswer: Let's think step by step.
UNIVERSAL MODE: You, as the moderator, will evaluate both sides' answers and determine if there is a clear
preference for an answer candidate. If so, please summarize your reasons for supporting affirmative/negative side and
give the final answer that you think is correct, and the debate will conclude. If not, the debate will continue to
the next round. Now please output your answer in json format, with the format as follows:
{\"Whether there is a preference\": \"Yes or No\", \"Supported Side\": \"Affirmative or Negative\",
\"Reason\": \"\", \"debate_answer\": \"the capital letter corresponding to the answer\"}.
Please strictly output in JSON format, do not output irrelevant content.

FINAL MODE: You, as the moderator, will evaluate both sides' answers and determine your
preference for an answer candidate. Please summarize your reasons for supporting affirmative/negative side and
give the final answer that you think is correct to conclude the debate. Now please output your answer in json format, with the format as follows:
{\"Supported Side\": \"Affirmative or Negative\", \"Reason\": \"\", \"debate_answer\": \"the capital letter corresponding to the answer\"}.
Please strictly output in JSON format, do not output irrelevant content.
```

ER CoT:
```
\n\nQuestion: {question}
\n\nExplanation: Let's solve this step-by-step, referring to authoritative sources as needed.
```

FEW SHOT:
```
\n\nQuestion: {question}
\n\nAnswer:
```

SIMPLE:
```
Instruction: Answer this multiple choice question.
\n\nInput: {question}
\n\nOutput: The Answer to the question is:
```

SPP:
```
SPP ORIGINAL:
When faced with a task, begin by identifying the participants who will contribute to solving the task. Provide profiles of the participants,
describing their expertise or needs. Then, initiate a multi-round collaboration process until a final solution is reached. The participants will
give critical comments and detailed suggestions whenever necessary.

    Here are some examples:
    ---
    Example Task 1: Use numbers and basic arithmetic operations (+ - * /) to obtain 24. You need to use all numbers, and each number can only be
    used once.
    Input: 6 12 1 1

    Participants: AI Assistant (you); Math Expert

    Profiles:
        - AI Assistant (you): A super-intelligent AI assistant capable of performing tasks more effectively than humans.
        - Math expert: A person who is good at math games, arithmetic calculation, and long-term planning.

    Start collaboration!

    Math Expert: Let's analyze the task in detail. You need to make sure that you meet the requirement, that you need to use exactly the four
    numbers (6 12 1 1) to construct 24. To reach 24, you can think of the common divisors of 24 such as 4, 6, 8, 3 and try to construct these
    first. Also you need to think of potential additions that can reach 24, such as 12 + 12.
    AI Assistant (you): Thanks for the hints! Here's one initial solution: (12 / (1 + 1)) * 6 = 24
    Math Expert: Let's check the answer step by step. (1+1) = 2, (12 / 2) = 6, 6 * 6 = 36 which is not 24! The answer is not correct. Can you
    fix this by considering other combinations? Please do not make similar mistakes.
    AI Assistant (you): Thanks for pointing out the mistake. Here is a revised solution considering 24 can also be reached by 3 * 8:
    (6 + 1 + 1) * (12 / 4) = 24.
    Math Expert: Let's first check if the calculation is correct. (6 + 1 + 1) = 8, 12 / 4 = 3, 8 * 3 = 24. The calculation is correct, but you
    used 6 1 1 12 4 which is not the same as the input 6 12 1 1. Can you avoid using a number that is not part of the input?
    AI Assistant (you): You are right, here is a revised solution considering 24 can be reached by 12 + 12 and without using any additional
    numbers: 6 * (1 - 1) + 12 = 24.
    Math Expert: Let's check the answer again. 1 - 1 = 0, 6 * 0 = 0, 0 + 12 = 12. I believe you are very close, here is a hint: try to change
    the "1 - 1" to "1 + 1".
    AI Assistant (you): Sure, here is the corrected answer:  6 * (1+1) + 12 = 24
    Math Expert: Let's verify the solution. 1 + 1 = 2, 6 * 2 = 12, 12 + 12 = 12. You used 1 1 6 12 which is identical to the input 6 12 1 1.
    Everything looks good!

    Finish collaboration!

    Final answer: 6 * (1 + 1) + 12 = 24

    ---
    Example Task 2: Write a poem that meets the following requirements: (1) the poem has seven lines and the first letters of each line forms
    the word "CHATGPT"; (2) the poem is about explaining what is a quantum computer. (3) the poem needs to be easy to understand by a ten
    years old kid.

    Participants: AI Assistant (you); Poet; Computer Scientist; Ten year old child

    Profiles:
        - AI Assistant (you): A super-intelligent AI assistant capable of performing tasks more effectively than humans.
        - Poet: A person who studies and creates poetry. The poet is familiar with the rules and formats of poetry and can provide guidance on
        how to write a poem.
        - Computer Scientist: A scholar who specializes in the academic study of computer science. The computer scientist is familiar with
        the concept of a quantum computer and can provide guidance on how to explain it.
```

```
    - Ten year old child: A child with a limited English vocabulary and little knowledge about complicated concepts, such as a quantum
    computer.

Poet: Make sure that you write the poem with seven lines, and the first letters of the lines should be C, H, A, T, G, P, T.
Computer Scientist: A quantum computer is an advanced computing device that uses the principles of quantum mechanics to process and store
information. Unlike classical computers that use bits to represent information as 0s and 1s, quantum computers use quantum bits or qubits.
Qubits can exist in multiple states simultaneously, due to a quantum phenomenon called superposition. You can consider using these
information for the poem.
Ten year old child: I hope the poem to be fun and easy to understanding. I don't want to see a lot of jargons or complicated concepts.
AI Assistant (you): Thanks for the guidance! Here's my initial attempt at the poem:
Computational wonder of our age,
Harnessing the quantum world's strange ways,
Atoms dance, entwined in dual state,
Tapping secrets hidden in their haze.

Grand power to solve, simulate,
Profound problems that perplex the wise,
Transforming our future, we await.

Poet: Let's verify if the poem meets the requirements. The first letters are CHATGPT which is correct! And the poem rhymes well. Good job!
Computer Scientist: Everything looks good to me!
Ten year old child: I don't know what does perplex mean. Can you make the use of words easier to understand?
AI Assistant (you): Sure, let me revise the poem by using more common words. Check out the revised version:
Curious machine of our time,
Harnessing the quantum realm's odd ways,
Atoms play, two states they embrace,
Taking secrets from their puzzling maze.

Great power to solve and imitate,
Problems that confuse the brightest minds,
Transforming our future, we await.

Poet: Let's check again if the poem meets the requirements. The first letters are C H A T G P T. And now the poem is more accessible to
children. Everything looks good to me.
Computer Scientist: Looking good!
Ten year old child: I like this version a lot!

Finish collaboration!

Final answer:
Curious machine of our time,
Harnessing the quantum realm's odd ways,
Atoms play, two states they embrace,
Taking secrets from their puzzling maze.

Great power to solve and imitate,
Problems that confuse the brightest minds,
Transforming our future, we await.

---
Now, identify the participants, provide their profiles, and collaboratively solve the following task step by step. Remember to provide the
final solution with the following format "Final answer: (a single capital letter).".

Task: Answer this multiple choice question: \n\nInput: {question}
SPP EXPERT:
When faced with a task, begin by identifying the participants who will contribute to solving the task. Note that the participants can
only be
either AI Assistant (you) or Expert. Then, initiate a multi-round collaboration process until a final conclusion is reached.  The Expert will
give critical comments and detailed suggestions whenever necessary.

Here are some examples:
---
Example Task 1: Use numbers and basic arithmetic operations (+ - * /) to obtain 24. You need to use all numbers, and each number can only
be used once.
Input: 6 12 1 1

Participants: AI Assistant (you); Expert

Start collaboration!

Expert: Let's analyze the task in detail. You need to make sure that you meet the requirement, that you need to use exactly the four
numbers (6 12 1 1) to construct 24. To reach 24, you can think of the common divisors of 24 such as 4, 6, 8, 3 and try to construct these
first. Also you need to think of potential additions that can reach 24, such as 12 + 12.
AI Assistant (you): Thanks for the hints! Here's one initial solution: (12 / (1 + 1)) * 6 = 24
Expert: Let's check the answer step by step. (1+1) = 2, (12 / 2) = 6, 6 * 6 = 36 which is not 24! The answer is not correct. Can you fix
this by considering other combinations? Please do not make similar mistakes.
AI Assistant (you): Thanks for pointing out the mistake. Here is a revised solution considering 24 can also be reached by 3 * 8:
(6 + 1 + 1) * (12 / 4) = 24.
Expert: Let's first check if the calculation is correct. (6 + 1 + 1) = 8, 12 / 4 = 3, 8 * 3 = 24. The calculation is correct, but you used
6 1 1 12 4 which is not the same as the input 6 12 1 1. Can you avoid using a number that is not part of the input?
AI Assistant (you): You are right, here is a revised solution considering 24 can be reached by 12 + 12 and without using any additional
numbers: 6 * (1 - 1) + 12 = 24.
Expert: Let's check the answer again. 1 - 1 = 0, 6 * 0 = 0, 0 + 12 = 12. I believe you are very close, here is a hint: try to change the
"1 - 1" to "1 + 1".
AI Assistant (you): Sure, here is the corrected answer:  6 * (1+1) + 12 = 24
Expert: Let's verify the solution. 1 + 1 = 2, 6 * 2 = 12, 12 + 12 = 12. You used 1 1 6 12 which is identical to the input 6 12 1 1.
Everything looks good!

Finish collaboration!

Final answer: 6 * (1 + 1) + 12 = 24
```

```
    ---
    Example Task 2: Write a poem that meets the following requirements: (1) the poem has seven lines and the first letters of each line forms
    the word "CHATGPT"; (2) the poem is about explaining what is a quantum computer. (3) the poem needs to be easy to understand by a ten years
    old kid.

    Participants: AI Assistant (you); Expert

    Expert: Make sure that you write the poem with seven lines, and the first letters of the lines should be C, H, A, T, G, P, T. A quantum
    computer is an advanced computing device that uses the principles of quantum mechanics to process and store information. Unlike classical
    computers that use bits to represent information as 0s and 1s, quantum computers use quantum bits or qubits. Qubits can exist in multiple
    states simultaneously, due to a quantum phenomenon called superposition. You can consider using these information for the poem. I hope the
    poem to be fun and easy to understanding. I don't want to see a lot of jargons or complicated concepts.
    AI Assistant (you): Thanks for the guidance! Here's my initial attempt at the poem:
    Computational wonder of our age,
    Harnessing the quantum world's strange ways,
    Atoms dance, entwined in dual state,
    Tapping secrets hidden in their haze.

    Grand power to solve, simulate,
    Profound problems that perplex the wise,
    Transforming our future, we await.

    Expert: Let's verify if the poem meets the requirements. The first letters are CHATGPT which is correct! And the poem rhymes well. Good
    job! I don't know what does perplex mean. Can you make the use of words easier to understand?
    AI Assistant (you): Sure, let me revise the poem by using more common words. Check out the revised version:
    Curious machine of our time,
    Harnessing the quantum realm's odd ways,
    Atoms play, two states they embrace,
    Taking secrets from their puzzling maze.

    Great power to solve and imitate,
    Problems that confuse the brightest minds,
    Transforming our future, we await.

    Expert: Let's check again if the poem meets the requirements. The first letters are C H A T G P T. And now the poem is more accessible
    to children. Everything looks good to me. I like this version a lot!

    Finish collaboration!

    Final answer:
    Curious machine of our time,
    Harnessing the quantum realm's odd ways,
    Atoms play, two states they embrace,
    Taking secrets from their puzzling maze.

    Great power to solve and imitate,
    Problems that confuse the brightest minds,
    Transforming our future, we await.

    ---

    Now, identify the participants and collaboratively solve the following task step by step. Note that the participants can only be either
    AI Assistant (you) or Expert. Remember to provide the final solution with the following format "Final answer: (a single capital letter).

    Task: Answer this multiple choice question: \n\nInput: {question}
SPP JUDGE:
Instruction: You serve as the moderator in this debate. At each opportunity
you will critic the responses of each of the agents and guide the conversation.
You will then make a clear decision by providing the most likely capital letter
answer at the end.
\n\nInput: {question}
\n\nAnswer:
```