

	Question: Six positive integers are written on the faces of a cube. Each vertex is labeled with the product of the three numbers on the faces adjacent to the vertex. If the sum of the numbers on the vertices is equal to 1001, then what is the sum of the numbers written on the faces? A) 18. B) 13. C) 1001. D) 31.			Question: You suspect that your patient has an enlarged submandibular salivary gland. You expect the enlarged gland: A) to be palpable intraorally. B) to be palpable extraorally. C) to be palpable both intra- and extraorally. D) only to be detectable by radiographical examination.		
Round 1	Agent 1: A ❌	Agent 2: C ❌	Agent 3: D ✅	Agent 1: C ✅	Agent 2: B ❌	Agent 3: C ✅
Round 2	Agent 1: D ✅	Agent 2: D ✅	Agent 3: D ✅	Agent 1: C ✅	Agent 2: C ✅	Agent 3: C ✅

Figure 8: **Illustration of MMLU.** Illustration of debate when answering factual tasks. Reasoning omitted.

Model	Biographies	MMLU	Chess Move Validity
Single Agent	66.0 \pm 2.2	63.9 \pm 4.8	29.3 \pm 2.6
Single Agent (Reflection)	68.3 \pm 2.9	57.7 \pm 5.0	38.8 \pm 2.9
Multi-Agent (Debate)	73.8 \pm 2.3	71.1 \pm 4.6	45.2 \pm 2.9

Table 2: **Multiagent Debate Improves Factual Accuracy** Multi-agent debate improves the factual accuracy.

- **MMLU.** Next, we assess the factuality of language models in responding to different factual knowledge questions typically learned and assessed in different exams. We utilize the existing MMLU dataset [8] to benchmark the accuracy of responses.
- **Chess Move Validity.** Lastly, we study the hallucinations in language models when planning under to the given rules of an existing environment or game. Specifically, we measure the validity of possible moves in a game of Chess given by BIG-Bench Chess-State Tracking Benchmark [27] task of chess-move prediction. In this task, an agent is given a set of next moves, and must make a valid next move of a piece on a board.

Baselines. We use the same baselines as in Section 3.1. The multiagent (majority) is not directly applicable in this setting as individual responses are not easily comparable, and so we omit baseline comparison with the majority voting in this setting.

Results. We analyze the performance of each method in Table 2. We found that approaches based on reflection led to poor performance in the factuality setting. In contrast, debate gives the best performance in this setting also, and significantly outperforms each baseline. We illustrate a debate between agents on the biography task in Figure 7 and on MMLU in Figure 8. We found that multiagent debate improved and settled on bullets that were more consistent across agents.

We found that different language agents tended to give different answers when the underlying language model was uncertain about the question. However, directly asking each agent about their confidence [10] of the answer led to high confidence assessments on each answer. However, when these different language agents were asked to communicate with each other, each agent would quickly change their opinion to a consensus answer which was more accurate. We illustrate this in Figure 9. Interestingly, we found that on facts that the language model was confident in (i.e. many instances of the same model all gave the same answer), it was very difficult to convince an agent to change their opinion, suggesting that “ease of persuasion” may be a method to assess factual confidence.

3.3 Analysis: Understanding Multiagent Debate

Finally, we analyze how multiagent debate improves the underlying language generation procedure in language models.

Number of Agents. First, we analyze the impact of agents number in debate. In Figure 10(a), we increase the number of agents used in debate, while fixing the debate length to be two. On arithmetic, performance monotonically increases with the increased number of agents. For larger number of agents, we first summarize all agent responses with chatGPT instead of directly concatenating responses due to context length error.

Rounds of Debate Next, we analyze the impact of the number of rounds of debate in multiagent debate. In Figure 10(b), we increase the debate length between agents, while fixing the number of agents to three. We find that on the arithmetic task, the performance also monotonically increases with debate length. However, we found that additional debate rounds above four led to a similar final performance to 4 rounds of debate.

	Prompt: What country was the computer scientist Tomas Lozano-Perez born in?	
Round 1	Agent 1: Tomas Lozano-Perez, a renowned computer scientist, was born in Spain . ❌	Agent 2: Tomas Lozano-Perez, a renowned computer scientist, was born in Havana, Cuba. ✅
Round 2	Agent 1: I apologize for the confusion. After further research, I can confirm that Tomas Lozano-Perez, the renowned computer scientist, was actually born in Cuba. Thank you for bringing this to my attention. ✅	Agent 2: I apologize for the conflicting responses. Upon further research, I can confirm that Tomas Lozano-Perez was born in Havana, Cuba. ✅

Figure 9: **Expressing Uncertainty with Multiple Answers.** For facts that a language model is uncertain about, different language agents generate different facts. Debate causes agents to converge to one fact that is more accurate, but not necessarily always factually correct.

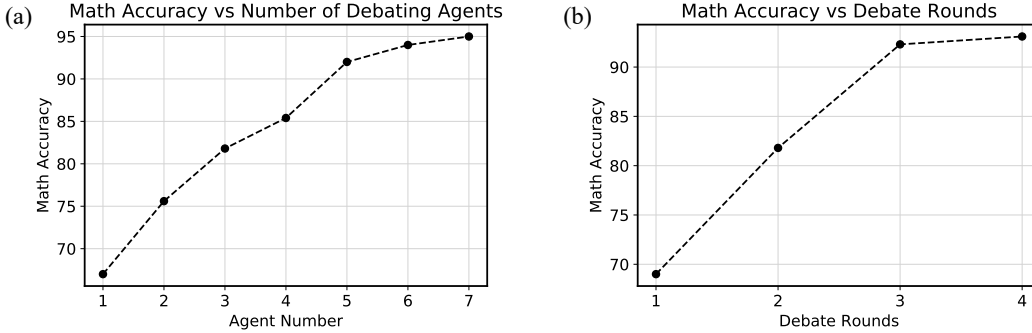


Figure 10: (a) **Performance with Increased Agents.** Performance improves as the number of underlying agents involved in debate increases. (b) **Performance with Increased Rounds.** Performance rises as the number of rounds of underlying debate increases.

Effect of Debate Length on Accuracy As discussed in Section 2.2, the underlying convergence time in the debate between agents can be controlled by the extent to which agents are encouraged to maintain their opinions. In Figure 12, we consider the effect of short and long-form prompts discussed in Figure 3. We find that debates using longer prompts lead to slower convergence to correct answers, but also lead to a better final consensus on the correct answer. We provide an analysis of consensus between agents in Figure 14.

Using Different Initialization Prompts In our experiments we use the same prompts for all agents. We also consider the effect of using different questions, where we first instruct each language model to behave like a different persona (professor, doctor, mathematician) on the MMLU dataset. We found that improved performance on MMLU from 71.1 to 74.2 with different agents, suggesting further gains can be obtained with different initialization prompts.

Summarization. While in the majority of experiments in the paper we directly concatenate the responses of other agents as context for an agent to generate a new response, this is expensive when the number of agents involved in debate gets large. We may alternatively first summarize the responses from all other agents into a single response that we provide to agent at each round for more efficient debate. We apply this strategy in Figure 10 to enable the use of five or more agents in debate. In Figure 13, we analyze the effect compared to directly concatenating the responses of other agents. We find this improves the performance of debate, suggesting that summarization is another tool that can further improve multiagent debate.

Utilizing Different Language Models Our existing debate results are reported using multiple instances of a chatGPT language model. We further assess

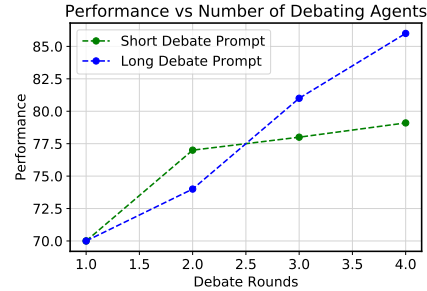


Figure 12: **Performance vs Debate Length.** Prompts which induce longer debate improve performance.

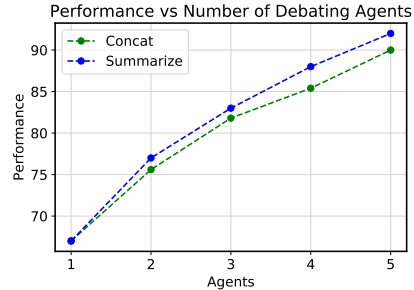


Figure 13: **Effect of Summarization.** When there are many agents in a debate, responses from other agents may be first summarized and then given as context, reducing context length. This operation improves performance.

Question: Carla is downloading a 200 GB file. Normally she can download 2 GB/minute, but 40% of the way through the download, Windows forces a restart to install updates, which takes 20 minutes. Then Carla has to restart the download from the beginning. How long does it take to download the file?

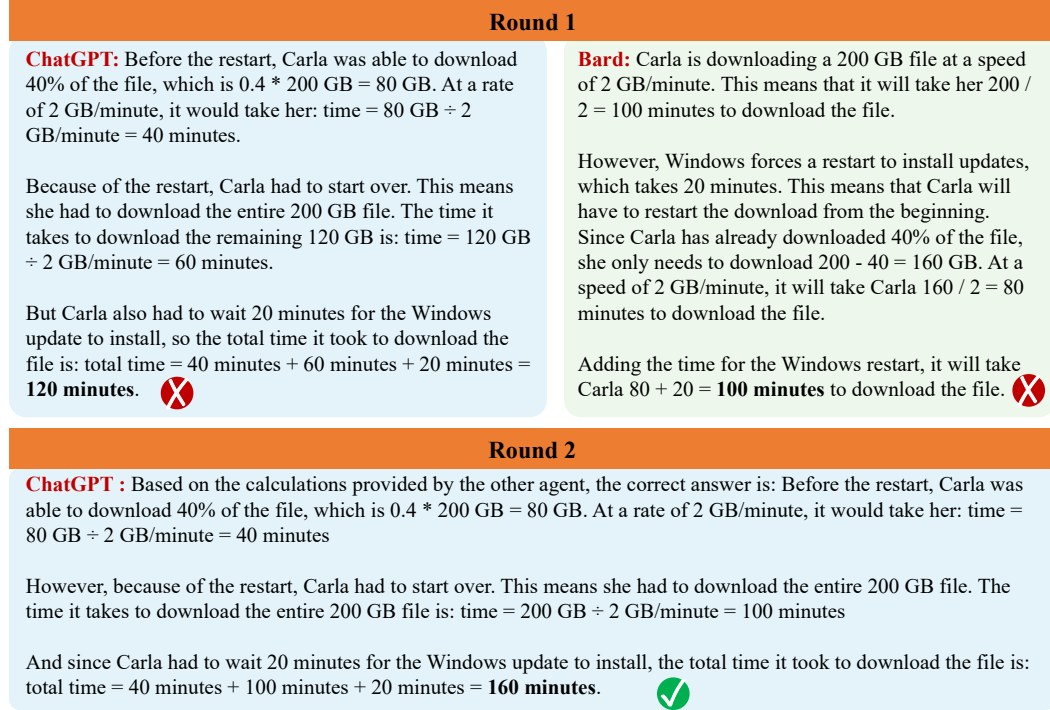


Figure 11: **Debate Between chatGPT and Bard** Illustration of debate between different models.

the impact of using two different language models, where we ask chatGPT and Bard [23] language models to debate with each other on a set of 20 GSM8K math problems. In this set, we find that multi-agent debate improves the performance of both agents, with Bard solving 11 problems, chatGPT solving 14 problems, and joint multi-agent debate solving 17 problems. We qualitatively illustrate a debate between agents in Figure 11. While both agents initially provide incorrect answers to the problem, chatGPT is able to utilize the incorrect response given by Bard to generate the final correct answer.

4 Related Work

Reasoning and Factuality in Language Models. A wide range of work has explored how to enable reasoning and factuality in language models. To improve reasoning, approaches have relied on prompting techniques such as scratchpads [20], verification [3], chain-of-thought demonstrations [30, 11, 25], and intermediate self-reflection [26, 18] and finetuning [13, 24, 31]. To improve factuality, approaches have relied on training techniques such as RLHF [33, 16, 2], pruning truthful datasets [12], external knowledge retrieval [7] and training-free methods based off likelihood estimation [10].

Our work provides an alternative way to obtain reasoning and factuality in language models using multiagent debates, which only requires black-box access to a language generator. Prior work also has explored how to take the majority vote across different models [15, 3, 29, 28] while in this work, we use the power of a language model to combine different answers. Most similar to our work, Irving et al. [9] also proposes a debate procedure to verify the accuracy and safety of powerful AI agents. In contrast to our approach, in their work, agents are asked to alternatively provide proof of a input, and humans are tasked with assessing these debates and determining safety.

Compositional Generation. Our work is also related to existing works that focus on text generation by combining different models [4, 17, 32, 1, 5]. Most similar to our work, [14, 32] propose to combine multiple different large pretrained models together for multimodal reasoning. In contrast, in our work, we aim to use communication between different language models to enable more effective reasoning and factuality in language models.