

Improving Factuality and Reasoning in Language Models through Multiagent Debate

Yilun Du
MIT CSAIL
yilundu@mit.edu

Shuang Li
MIT CSAIL
lishuang@mit.edu

Antonio Torralba
MIT CSAIL
torralba@mit.edu

Joshua B. Tenenbaum
MIT CSAIL, BCS, CBMM
jbt@mit.edu

Igor Mordatch
Google Brain
imordatch@google.com

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in language generation, understanding, and few-shot learning in recent years. An extensive body of work has explored how their performance may be further improved through the tools of prompting, ranging from verification, self-consistency, or intermediate scratchpads. In this paper, we present a complementary approach to improve language responses where multiple language model instances propose and debate their individual responses and reasoning processes over multiple rounds to arrive at a common final answer. Our findings indicate that this approach significantly enhances mathematical and strategic reasoning across a number of tasks. We also demonstrate that our approach improves the factual validity of generated content, reducing fallacious answers and hallucinations that contemporary models are prone to. Our approach may be directly applied to existing black-box models and uses identical procedure and prompts for all tasks we investigate. Overall, our findings suggest that such "society of minds" approach has the potential to significantly advance the capabilities of LLMs and pave the way for further breakthroughs in language generation and understanding. Project website at https://composable-models.github.io/llm_debate/.

1 Introduction

Large language models (LLMs) have demonstrated remarkable language generation, understanding, and few-shot learning capabilities in recent years. These methods are trained on a massive corpus of text on the internet, where the quality and accuracy of extracted natural language may not be ensured. Thus, current models may suffer from confidently hallucinating facts or making implausible jumps in chains of reasoning. An extensive body of recent work has focused on improving factual accuracy and reasoning in language models. These range from prompting models with few or zero-shot chain-of-thought demonstrations, use of verification, self-consistency, or intermediate scratchpads.

We note that these techniques are applied over a single model instance. Instead, we propose a complementary approach inspired by *The Society of Mind* [19] and multi-agent settings, where multiple language model instances (or agents) individually propose and jointly debate their responses and reasoning processes to arrive at a single common answer. More specifically, given a query, multiple instances of a language model first generate individual candidate answers to a query. Then each individual model instance reads and critiques the responses of all other models and uses this content to update its own answer. This step is then repeated over several rounds. This process induces models to construct answers that are consistent with both their internal critic as well as sensible in

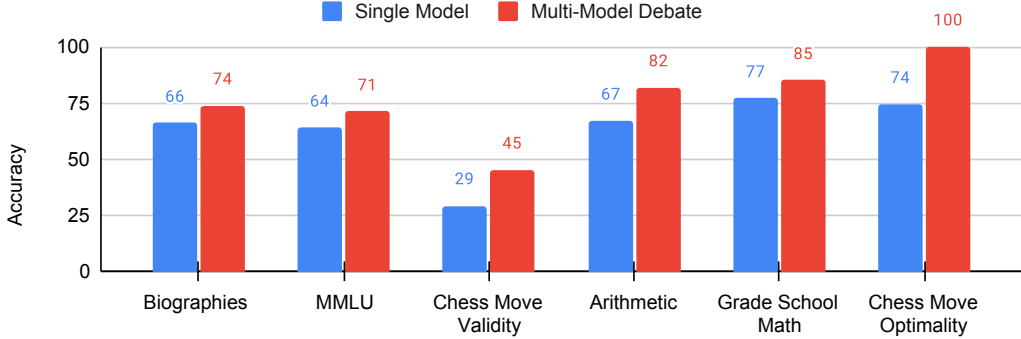


Figure 1: **Multiagent Debate Improves Reasoning and Factual Accuracy.** Accuracy of traditional inference and our multi-agent debate over six benchmarks (chess move optimality reported as a normalized score)

light of the responses of other agents. The resulting quorum of models can hold and maintain multiple chains of reasoning and possible answers simultaneously before proposing the final answer.

We find that our debate approach outperforms single model baselines such as zero-shot chain of thought [11] and reflection [26, 18] on a variety of six reasoning, factuality, and question-answering tasks. Using both multiple model agents and multiple rounds of debate are important to achieve the best performance. Given an initial query, we find that individual model instances propose a diverse range of answers despite being the same model class (although we also investigate the case of mixing different model types, such as chatGPT [21] and Bard [23]). After debating and examining the responses of other model instances, we find that the population almost always converges on a single and more accurate common answer. Debate results are also less likely to include false facts that models are internally uncertain of. This is because as the debate progresses, individual model instances tend to disagree on uncertain facts and omit them from the answer (Figure 7). Lastly, we find that debate does not just act to amplify one correct answer in a model quorum - we find many cases where all the models initially make incorrect predictions, but then arrive at the correct answer as debate progresses (Figure 4, 11).

We use the same methodology and prompt templates for all our tasks and require only black-box access to language model generations – no model-internal information such as likelihoods or gradients is needed. This allows our method to be used with common public models serving interfaces. The method is also orthogonal to other model generation improvements such as retrieval or prompt engineering (in fact, we combine our debate method with zero-shot chain of thought). While the debate process is more costly, requiring multiple model instances and rounds, it arrives at significantly improved answers and may be used to generate additional model training data, effectively creating a model self-improvement loop.

To help evaluate the effect of our approach on factual accuracy, we introduce a new benchmark and dataset evaluating factual accuracy of famous computer scientist biographies. We find that contemporary language models have an especially high tendency to hallucinate factually incorrect biographies, often misrepresenting the relevant institutions and dates. Moreover, these facts often inconsistent across different language model instances. By asking models to come to a consensus across their answers, such inconsistent facts may be either removed or corrected.

In summary, our work contributes the following. First, we present a novel approach to improving factual correctness and reasoning accuracy in contemporary language models, leveraging a multi-agent debate process between models. Second, we introduce a new benchmark of factual correctness which contemporary language models struggle with. Finally, we evaluate the performance of our debate procedure in language generation, both in terms of the number of agents, the underlying rounds of debate, and the prompts that elicit such behavior across a set of six different reasoning and factual accuracy tasks.

2 Language Generation through Multiagent Debate

We present an approach to generate language responses through multiagent debate. We provide an overview of our approach in Section 2.1. We further discuss convergence to consensus in the debate process in Section 2.2. The overall overview of our approach is shown in Figure 2.

User Input: A treasure hunter found a buried treasure chest filled with gems. There were 175 diamonds, 35 fewer rubies than diamonds, and twice the number of emeralds than the rubies. How many of the gems were there in the chest?

Round 1	
<p>Agent 1: Let x be the number of rubies. Then the number of diamonds is 175. And the number of emeralds is $2(x-35)$, since there are twice as many emeralds as rubies and there are 35 fewer rubies than diamonds.</p> <p>To find the total number of gems, we add the number of diamonds, rubies, and emeralds: Total gems = $175 + x + 2(x-35) = 3x + 105$.</p> <p>We don't know the exact value of x, but we do know that it represents the number of rubies, which is less than 175 (since there are 35 fewer rubies than diamonds). Let's try some values of x: If $x = 50$, then the total gems = $3(50) + 105 = 255$. If $x = 100$, then the total gems = $3(100) + 105 = 405$.</p> <p>Since we know there were fewer rubies than diamonds, the answer must be closer to 255 than to 405. After some trial and error, we can find that $x = 40$ gives us: Total gems = $3(40) + 105 = 225$. ❌</p>	<p>Agent 2: If there were 175 diamonds, then the number of rubies is 35 fewer, or $175 - 35 = 140$ rubies.</p> <p>And since there were twice as many emeralds as rubies, we can represent the number of emeralds as $2 * 140 = 280$ emeralds.</p> <p>To find the total number of gems in the chest, we can add up the number of diamonds, rubies, and emeralds: $175 + 140 + 280 = 595$. Therefore, there were 595 gems in the chest. ✅</p>
<p>User Input: Using the solutions from other agents as additional information, can you give an updated response.... (Copy Agent 2's reply) (Copy Agent 1's reply)</p>	
Round 2	
<p>Agent 1: Given the information provided in other agents' solutions, we have two answers: One agent found that the number of rubies is 40 and the total number of gems is 225. Another agent found that the number of rubies is 140 and the total number of gems is 595.</p> <p>We know that the number of rubies is 35 fewer than the number of diamonds, which is 175. So the number of rubies should be $175 - 35 = 140$, as the second agent found. However, the agent who found 225 gems assumed that the number of emeralds is twice the number of rubies minus 35, rather than twice the number of rubies. If we use this assumption, we get: Number of rubies = 140. Number of emeralds = $2 * 140 - 35 = 245$. Total number of gems = $175 + 140 + 245 = 560$. Our final answer is 560. ❌</p>	<p>Agent 2: After reviewing the solutions provided by other agents, I agree with the second agent who found that the number of rubies is 140 and the total number of gems in the chest is 595. Therefore, my answer to the math problem is 595. ✅</p>
<p>User Input: Using the solutions from other agents as additional information, can you give an updated response.... (Copy Agent 2's reply) (Copy Agent 1's reply)</p>	
Round 3	
<p>Agent 1: After reviewing the solutions provided by other agents, I agree with the second agent who found that the number of rubies is 140 and the total number of gems in the chest is 595. Therefore, my answer to the math problem is 595. ✅</p>	<p>Agent 2: After reviewing the solutions provided by other agents, I also agree that the number of rubies is 140 and the total number of gems in the chest is 595. Therefore, my answer to the math problem is also 595. ✅</p>

Figure 2: **Illustration of Debate.** Illustration of the debate procedure.

2.1 Multiagent Language Generation

Consider your work process when solving the following math question on an exam: “What is the area of a triangle with side lengths of 3, 4, 5?”. In one thread of work, you may recognize that the triangle side-lengths directly correspond to a right triangle, and thus directly compute the area as $0.5 \times 3 \times 4 = 6$. To make sure that you have the right answer, you may then try to solve the problem differently by estimating an angle θ in the triangle using the Law of Cosines, and then obtain the area by using the formula $0.5 \times 3 \times 4 \times \sin(\theta)$, arriving at another answer to the given exam problem.

When these lines of work give the same answer, your confidence about the answer increases. In contrast, when these answers are different, individual lines of work may engage in a mental “debate” procedure, where you closely cross-examine the reasoning and assumptions of each line of work and refine solutions until a consistent answer.

Similarly, consider writing a biography of a historical figure. To ensure the factuality of the biography, you may consult multiple different sources on each fact. Facts that are consistent in each source increase your confidence about the fact. In contrast, facts that are inconsistent require careful cross-examination between sources to determine the final consistent data.