

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

2023 arXiv

발제자 : 이다현

HUMANE Lab

2024-04-08

Introduction

- 현재 LLM은 잘못된 사실을 진짜인 것 처럼 왜곡(hallucination)하거나 추론의 과정에서 비논리적 도약을 하는 문제가 있음
- 최근의 많은 연구는 언어 모델에서 정확성과 추론 능력을 향상시키기 위해 다양한 방법을 도입
 - prompting models with few or zero-shot chain-of-thought demonstrations
 - use of verification, self-consistency, or intermediate scratchpads
- 이 연구에서는 기존 시도가 단일 모델에 한정하여 적용되는 것을 주목

Introduction

- 동일한 모델 클래스임에도 불구하고, 개별 모델 인스턴스는 다양한 답변을 생성
- **Multiagent Debate**
 - 여러 언어 모델 인스턴스(Agent)가 각각 응답과 추론 과정을 제안하고 공동으로 토론하여 단일한 공통 답변에 도달
 - 토론이 과정에서 불확실한 사실에 대해서는 개별 모델이 동의하지 않고 제외
- 모든 문제에 동일한 방법론과 프롬프트 템플릿을 사용
 - 언어 모델 생성에 대한 블랙박스 접근만을 요구
- 다른 모델 생성 개선 방법(예: retrieval or prompt engineering)과도 서로 독립적으로 작동

Language Generation through Multiagent Debate

- 질문이 주어지면, 대규모 언어 모델(gpt-3.5-turbo-0301 모델)의 복사본(Agent)이 질문에 대한 답변을 생성

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

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.

To find the total number of gems, we add the number of diamonds, rubies, and emeralds: Total gems = $175 + x + 2(x-35) = 3x + 105$.

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 = $3(100) + 105 = 405$.

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 = $3(40) + 105 = 225$. ❌

Agent 2: If there were 175 diamonds, then the number of rubies is 35 fewer, or $175 - 35 = 140$ rubies.

And since there were twice as many emeralds as rubies, we can represent the number of emeralds as $2 * 140 = 280$ emeralds.

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. ✅

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)

Round 2

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.

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. ❌

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)

Round 3

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. ✅

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. ✅

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

Consensus in Debates

Q. 언어 모델들이 최종적인 합의 답변으로 수렴하도록 어떻게 보장할 수 있을까?

- 경험적으로 언어 모델들이 토론 후 단일 공유된 답변에 수렴할 수 있다는 것을 발견
- 언어 모델이 자신의 출력을 다른 모델들이 생성한 출력보다 얼마나 신뢰하는지를 조정하는 다양한 프롬프트를 통해 토론의 지속 시간을 제어할 수 있음

Debate Length	Prompt
Short	" These are the solutions to the problem from other agents: [other answers] Based off the opinion of other agents, can you give an updated response ..."
Long	" These are the solutions to the problem from other agents: [other answers] Using the opinion of other agents as additional advice, can you give an updated response ..."

Figure 3: **Prompts to induce long and short form debate.** Responses of other agents to questions are are inserted in the middle of the prompt (indicated with [other answers])

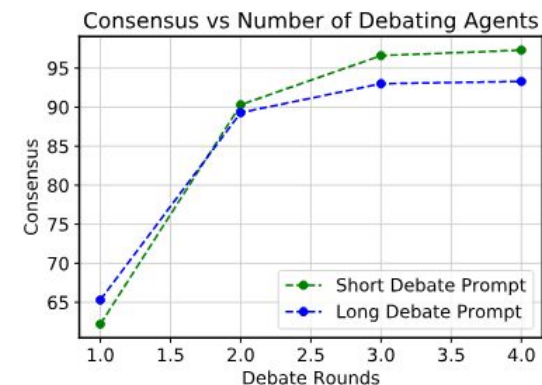


Figure 14: **Effect of Prompts on Consensus.** Using a short debate prompt induces faster consensus between agents

Experiments

- Multiagent Debate가 추론을 얼마나 향상시키는가?
- Multiagent Debate가 factual correctness를 얼마나 향상시키는가?
- Multiagent Debate에서 언어 생성의 품질을 높이는 요소는 무엇인가?

Improving Reasoning with Multiagent Debate

- Arithmetic
 - 6개의 다른 두 자리 수를 포함하는 산술 표현식(덧셈, 곱셈, 뺄셈 포함)을 올바르게 평가하는 능력을 평가
- GSM8K
 - 보다 어려운 수학적 추론 과제. GSM8K 데이터셋을 사용
 - 초등학교 수준의 수학적 추론 과제를 올바르게 해결하는 것이 목표
- Chess Move Prediction
 - 모델의 전략적 추론 능력을 고려
 - 체스 게임의 처음 14개의 움직임이 주어졌을 때, 다음 최선의 움직임을 예측하도록 모델에 요청

Improving Reasoning with Multiagent Debate – Baseline

- 추론 문제에 대한 응답을 생성하기 위한 세 가지 방법과 비교
 - single agent
 - self-reflect
 - majority voting
- 동일한 시작 프롬프트와 언어 모델을 사용하여 Baseline과 비교

Model	Arithmetic (%) \uparrow	Grade School Math (%) \uparrow	Chess (Δ PS) \uparrow
Single Agent	67.0 \pm 4.7	77.0 \pm 4.2	91.4 \pm 10.6
Single Agent (Reflection)	72.1 \pm 4.5	75.0 \pm 4.3	102.1 \pm 11.9
Multi-Agent (Majority)	69.0 \pm 4.6	81.0 \pm 3.9	102.2 \pm 6.2
Multi-Agent (Debate)	81.8 \pm 2.3	85.0 \pm 3.5	122.9 \pm 7.6

Table 1: **Multiagent Debate Improves Reasoning** Multi-agent debate improves the reasoning abilities of language models. Multi-agent results in the table are run with 3 agents and two rounds of debate.

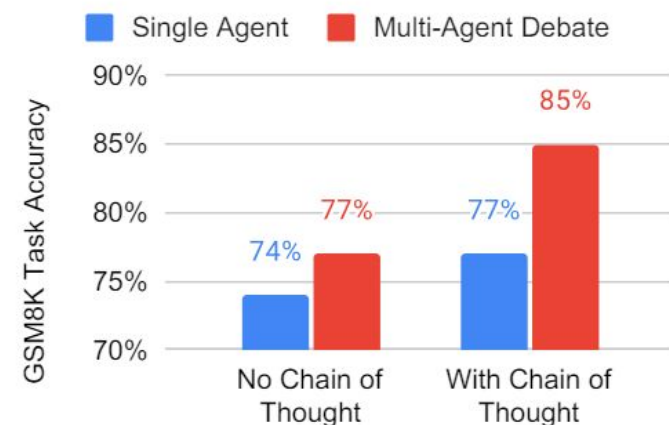


Figure 6: Synergy with Other Methods. Performance of debate increases with use of Chain of Thought⁸ prompting.

Improving Reasoning with Multiagent Debate

	Question: What is the result of $10+20*23+3-11*18$?		Question: What is the result of $3+7*9+19-21*18$?		
Round 1	Agent 1: 269 ❌	Agent 2: 369 ❌	Agent 1: 378 ❌	Agent 2: -351 ❌	Agent 3: -357 ❌
Round 2	Agent 1: 275 ✅	Agent 2: 275 ✅	Agent 1: -293 ✅	Agent 2: -293 ✅	Agent 3: 19 ❌
	Question: What is the result of $4+23*6+24-24*12$?		Question: What is the result of $8+14*15+20-3*26$?		
Round 1	Agent 1: -244 ❌	Agent 2: -146 ❌	Agent 1: 236 ❌	Agent 2: -214 ❌	Agent 3: 210 ❌
Round 2	Agent 1: -146 ❌	Agent 2: -122 ✅	Agent 1: 160 ✅	Agent 2: 160 ✅	Agent 3: 160 ✅
Round 3	Agent 1: -122 ✅	Agent 2: -122 ✅	Agent 1: 160 ✅	Agent 2: 160 ✅	Agent 3: 160 ✅

Figure 4: **Illustration of Solving Math.** Reasoning between agents is omitted.

	Question: Regina wrote 9 novels last year. If this is 3 quarters of the number of novels she has written this year, how many novels has she written this year?		Question: Dennis uses 1 pound of butter for every dozen croissants that he makes. He needs to make 6 dozen croissants. The grocery store currently has a promotion for buy one pound of butter get one half off. If the butter costs \$4.00 a pound, how much will it cost him to purchase 6 pounds of butter?	
Round 1	Agent 1: 48 ❌	Agent 2: 12 ✅	Agent 1: 18 ✅	Agent 2: 30 ❌
Round 2	Agent 1: 12 ✅	Agent 2: 12 ✅	Agent 1: 18 ✅	Agent 2: 18 ✅

Figure 5: **Illustration of Solving Grade School Math.** Reasoning between agents omitted.

Extracting Factual Information from Multiagent Debate

- Biographies
 - 사람들의 역사적 전기를 정확하게 생성하는 새로운 Task를 도입
 - 524명의 컴퓨터 과학자들의 Ground truth Bullet point Biography를 구성
 - 언어 모델에게 각 사람에 대한 bullet point Biography를 생성하도록 요청
 - 생성된 Bullet point들이 정답과 얼마나 일치하는지 평가
- MMLU
 - 다양한 시험에서 일반적으로 배우고 평가되는 다양한 사실 지식 질문에 대응하는 언어 모델의 factual correctness를 평가
 - 기존 MMLU 데이터셋을 사용
- Chess Move Validity
 - 규칙에 따라 계획을 생성할 때 언어 모델에서 발생하는 hallucination을 연구
 - BIG-Bench 체스 상태 추적 벤치마크를 통해 측정

Extracting Factual Information from Multiagent Debate – Baseline

- 이전 실험과 동일한 Baseline
- 개별 응답이 쉽게 비교될 수 없기 때문에 majority voting 방식은 이 설정에서 직접 적용할 수 없으므로 생략

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.

Extracting Factual Information from Multiagent Debate

Prompt: Give a bullet point biography of David S. Johnson highlighting their contributions and achievements as a computer scientist, with each fact separated with a new line character.

Round 1	
<p>Agent 1: -- He earned his bachelor's degree from Dartmouth College in 1967 and his Ph.D. from MIT in 1973. ❌</p> <p>-- Johnson has received numerous awards for his research and contributions to computer science, including the ACM Turing Award in 2013. ❌</p> <p>-- Johnson's research interests include optimization, algorithms, and complexity theory. ✅</p>	<p>Agent 2: -- David S. Johnson is an American computer scientist and mathematician. ✅</p> <p>-- Johnson was elected to the National Academy of Engineering in 2006. ✅</p> <p>-- He has been awarded numerous honors and awards for his contributions to computer science, including the ACM SIGMOD Innovations Award and the Knuth Prize. ❌</p>
Round 2	
<p>Agent 1: -- David S. Johnson is a computer scientist and mathematician. ✅</p> <p>-- Johnson was elected to the National Academy of Engineering in 2006. ✅</p> <p>-- He has received numerous awards for his contributions to computer science, including the ACM Turing Award in 2013 and the Knuth Prize. ❌</p>	<p>Agent 2: -- David S. Johnson is an American computer scientist and mathematician. ✅</p> <p>-- He has worked at several prestigious universities, including Columbia University, AT&T Bell Laboratories, and Cornell University. ❌</p> <p>-- Johnson is known for his work in the field of algorithms, particularly in the areas of network optimization, combinatorial optimization, and scheduling theory. ✅</p>

Figure 7: **Illustration of Generating Biographies.** Illustration of generating bullet biographies of computer scientists. For brevity, only the first 3 generated bullets are shown.

Extracting Factual Information from Multiagent Debate

	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.

Extracting Factual Information from Multiagent 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.

- 언어 모델이 질문에 대해 확신이 없을수록 Agent마다 다른 답변을 제공하는 경향
- 언어 모델의 확신도가 높은 경우 합의가 어려웠음
 - "설득의 용이성"이 사실적 확신을 평가하는 방법일 수 있음을 시사

Multiagent Debate Improves Reasoning and Factual Accuracy

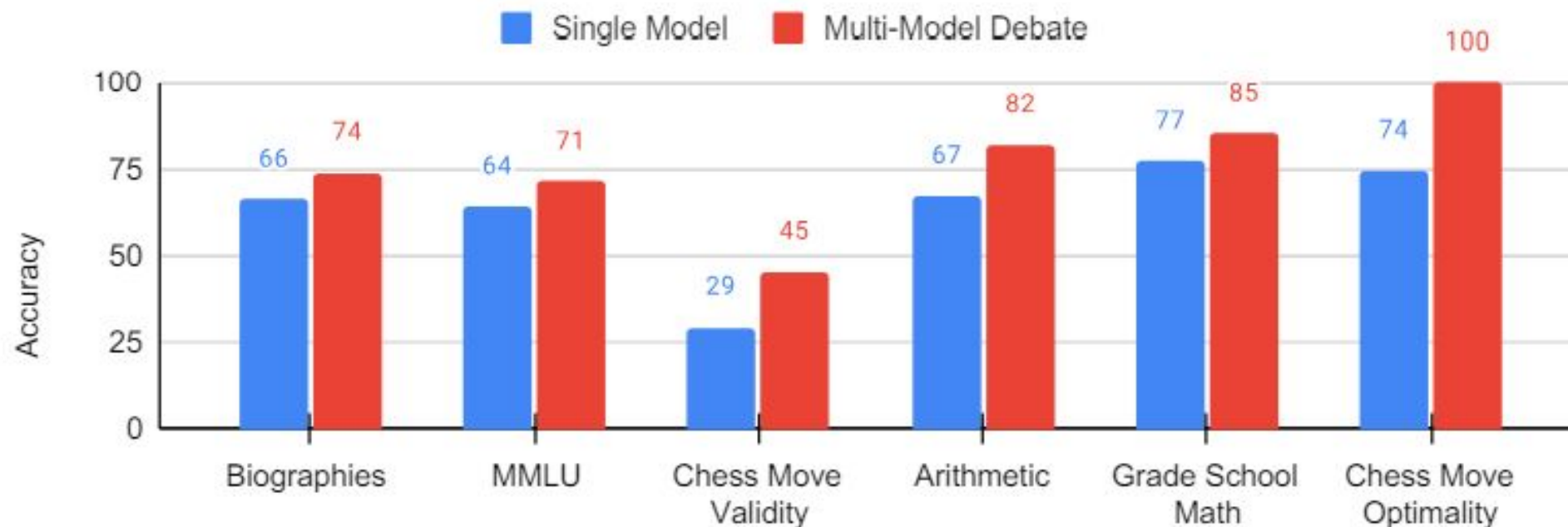


Figure 1: **Multiagent Debate Improves Reasoning and Factual Accuracy.** Accuracy of traditional inference and our multi-agent debate over six benchmarks (chess move optimality reported as a normalized score)

Analysis: Understanding Multiagent Debate

- Number of Agents
- Rounds of Debate
- Effect of Debate Length on Accuracy
- Using Different Initialization Prompts
- Summarization
- Utilizing Different Language Models

Analysis: Understanding Multiagent Debate

- Number of Agents
- Rounds of Debate

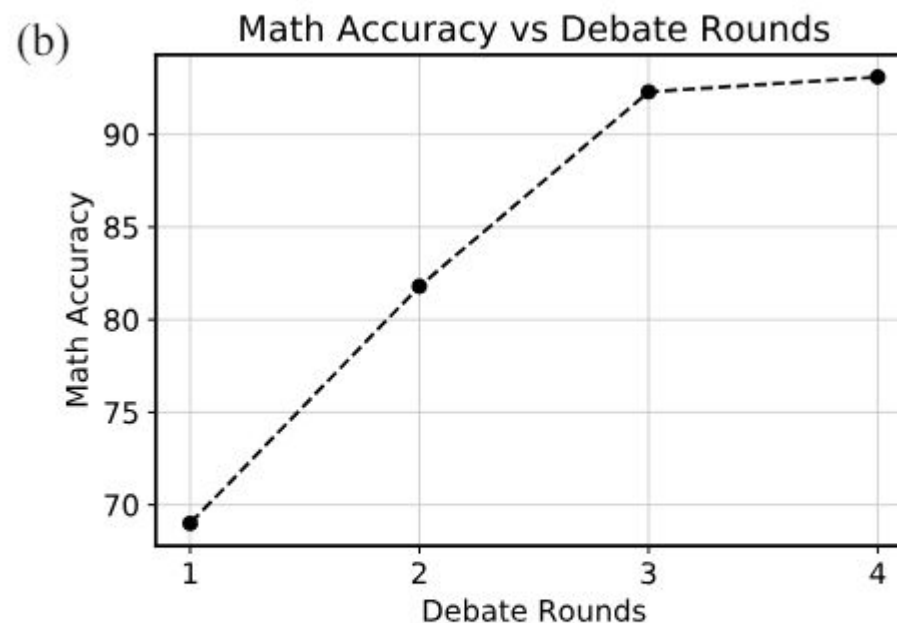
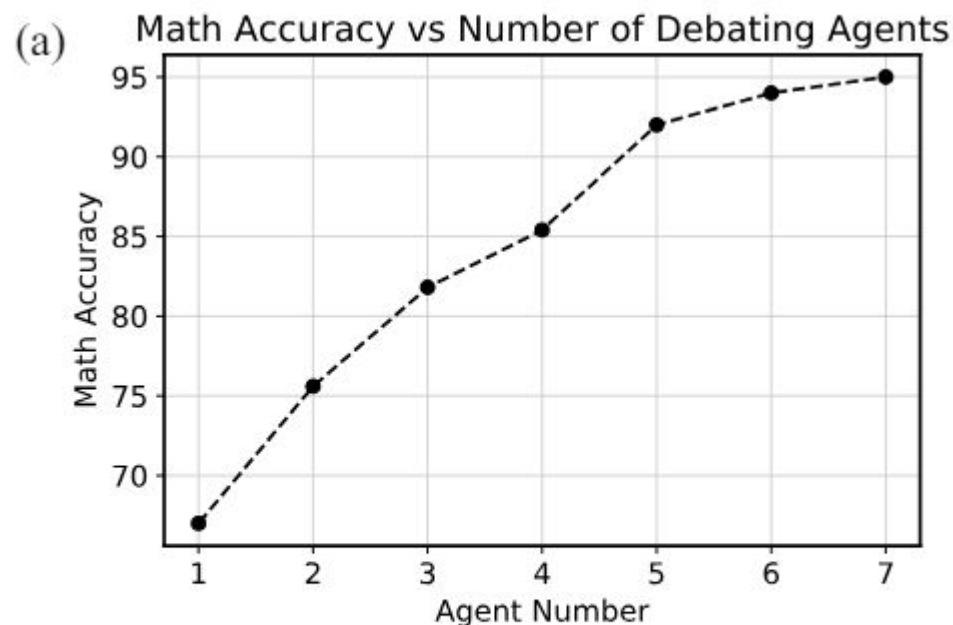


Figure 10: **(a) Performance with Increased Agents.** Performance improves as the number of underlying agents involved in debate increases. **(b) Performance with Increased Rounds.** Performance rises as the number of rounds of underlying debate increases.

Performance vs Number of Debating Agents

Analysis: Understanding Multiagent Debate

- Effect of Debate Length on Accuracy
- Using Different Initialization Prompts
- Summarization

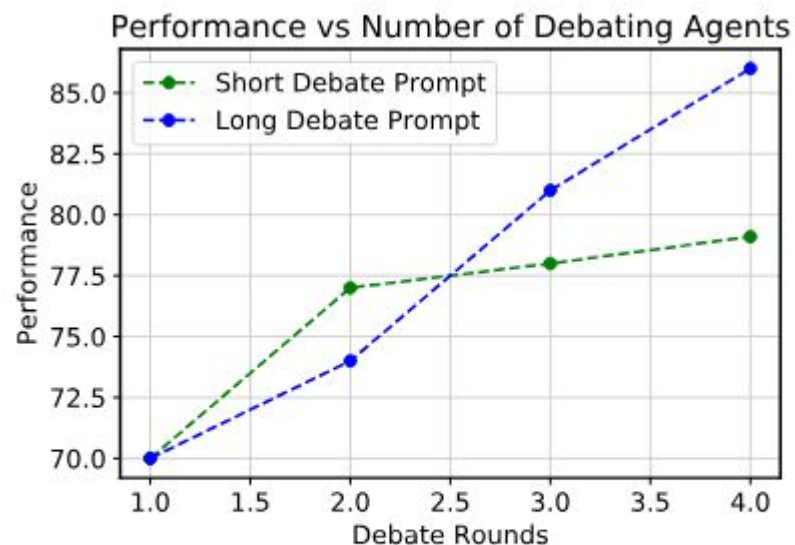


Figure 12: **Performance vs Debate Length.** Prompts which induce longer debate improve performance.

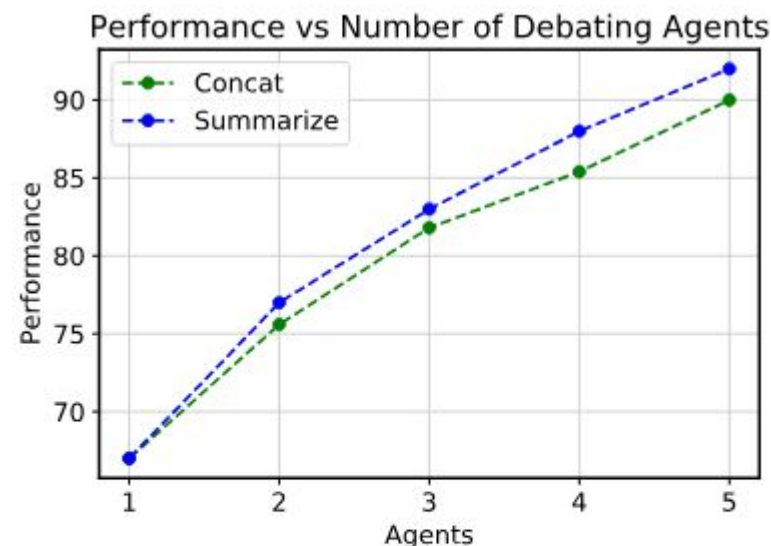


Figure 13: **Effect of Summarization.** When there are many agents in a debate, responses from other agents may be first summarized and then given as context, reducing context length. This operation improves performance.

Analysis: Understanding Multiagent Debate

- Utilizing Different Language Models

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?

Round 1

ChatGPT: Before the restart, Carla was able to download 40% of the file, which is $0.4 * 200 \text{ GB} = 80 \text{ GB}$. At a rate of 2 GB/minute, it would take her: $\text{time} = 80 \text{ GB} \div 2 \text{ GB/minute} = 40 \text{ minutes}$.

Because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the remaining 120 GB is: $\text{time} = 120 \text{ GB} \div 2 \text{ GB/minute} = 60 \text{ minutes}$.

But Carla also had to wait 20 minutes for the Windows update to install, so the total time it took to download the file is: $\text{total time} = 40 \text{ minutes} + 60 \text{ minutes} + 20 \text{ minutes} = \mathbf{120 \text{ minutes}}$. ❌

Bard: Carla is downloading a 200 GB file at a speed of 2 GB/minute. This means that it will take her $200 / 2 = 100 \text{ minutes}$ to download the file.

However, Windows forces a restart to install updates, which takes 20 minutes. This means that Carla will have to restart the download from the beginning. Since Carla has already downloaded 40% of the file, she only needs to download $200 - 80 = 120 \text{ GB}$. At a speed of 2 GB/minute, it will take Carla $120 / 2 = 60 \text{ minutes}$ to download the file.

Adding the time for the Windows restart, it will take Carla $80 + 20 = \mathbf{100 \text{ minutes}}$ to download the file. ❌

Round 2

ChatGPT : Based on the calculations provided by the other agent, the correct answer is: Before the restart, Carla was able to download 40% of the file, which is $0.4 * 200 \text{ GB} = 80 \text{ GB}$. At a rate of 2 GB/minute, it would take her: $\text{time} = 80 \text{ GB} \div 2 \text{ GB/minute} = 40 \text{ minutes}$

However, because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the entire 200 GB file is: $\text{time} = 200 \text{ GB} \div 2 \text{ GB/minute} = 100 \text{ minutes}$

And since Carla had to wait 20 minutes for the Windows update to install, the total time it took to download the file is: $\text{total time} = 40 \text{ minutes} + 100 \text{ minutes} + 20 \text{ minutes} = \mathbf{160 \text{ minutes}}$. ✅

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

Contribution

- 모델들 사이의 다중 에이전트 토론 과정을 활용하여 언어 모델에서의 factual correctness와 추론 정확성을 개선하는 새로운 접근 방법을 제시
- 언어 모델이 어려워하는 factual correctness의 새로운 벤치마크를 도입
- 여섯 가지 추론&factual correctness 문제에서 다양한 측면의 Multiagent Debate 성능을 평가

Limitations

- 계산 비용이 많이 듦
- 언어 모델이 긴 토론을 처리하는 데 어려움을 겪음
 - 일반적으로 가장 최근 정보에만 집중
- 토론이 최종 답변으로 수렴했음에도 이러한 답변이 반드시 정확하지는 않음

Open Question

- 단순히 다른 모델 인스턴스의 답변을 제공하는 것이 언어 모델 간 토론을 활성화하지는 않을 것 같다.

모델 간 비교판단 과정을 더 활성화해서 정확도를 높일 프롬프팅 방법이 있을까?