



**KOREA**  
UNIVERSITY

# 2025 동계세미나 발표

---

어수경

# Paper

- **Improving Factuality and Reasoning in Language Models through Multiagent Debate (ICML2024)**
- **Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? (ACL2024)**

# Paper

---

## **Improving Factuality and Reasoning in Language Models through Multiagent Debate**

---

**Yilun Du<sup>1</sup> Shuang Li<sup>1</sup> Antonio Torralba<sup>1</sup> Joshua B. Tenenbaum<sup>1</sup> Igor Mordatch<sup>2</sup>**

# AI Agents

- 최근 AI Agent 시대로의 전환 트렌드
- Planning에 기반해 reasoning chain generation, subtask decomposition → 각 agent가 tool, web browser 등 사용하여 수행
- **Mutiagent 활용 연구**
  - 주로 multiple agents를 활용함으로써 문제 해결
  - Complexity가 높은 task를 수행하기 위한 직관적인 방법으로 제시 (AgentVerse, ICLR 2024)
  - External feedback 기반의 수행 능력 향상 (MAD, EMNLP 2024)
  - 모듈화 및 상호보완적 결합 (AutoGen, Arxiv)
  - Self-reflection의 경우 degeneration-of-thoughts 문제 발생 : 스스로의 응답에 지나치게 confident한 경우 여러 라운드의 피드백을 거치더라도 새로운 사고 불가능 (Reconcile, ACL 2024)(MAD, EMNLP 2024)
  - Society-of-mind (Factuality~, ICML 2024)

# AI Agents

- 최근 AI Agent 시대로의 전환 트렌드
- Planning에 기반해 reasoning chain generation, subtask decomposition → 각 agent가 tool, web browser 등 사용하여 수행

- **Mutiagent 활용 연구**

- 주로 multiple agents를 활용함으로써 문제 해결
- Complexity가 높은 task를 수행하기 위한 직관적인 방법으로 제시 (AgentVerse, ICLR 2024)
- External feedback 기반의 수행 능력 향상 (MAD, EMNLP 2024)
- 모듈화 및 상호보완적 결합 (AutoGen, Arxiv)
- Self-reflection의 경우 degeneration-of-thoughts 문제 발생 : 스스로의 응답에 지나치게 confident한 경우 여러 라운드의 피드백을 거치더라도 새로운 사고 불가능 (Reconcile, ACL 2024)(MAD, EMNLP 2024)
- Society-of-mind (Factuality~, ICML 2024)

→ Multiagent collaboration 연구에 대해 파악/장점 확인

→ 문제점 발견 (궁극적으로는 AI agent 시대에 연구적으로 고려해야 할 사항들 파악)

# Introduction

- LLM의 발전에도 불구하고 hallucination, reasoning chain 내에서의 추론 건너뛰기 등 발생
- Factual accuracy, reasoning 능력을 향상시키는 연구들 → single model에 기반
- 본 연구의 특징: society of mind에 기반 (마음은 에이전트 (agents) 라고 하는 독특하고 간단한 프로세스 (process) 들이 거대한 조직을 이루어 그들이 상호작용을 한 결과물, Marvin Minsky)
- 여러 LLM들을 활용해서 각각 response를 제시, debate를 바탕으로 common answer에 도달
  - 주어진 query에 대해 multiple LLM들이 response generation
  - 각 LLM들은 다른 LLM들의 응답을 바탕으로 자신의 응답을 update

# Language Generation through Multiagent Debate

## [Expectation]

- Same answer → confidence increases
- Inconsistent answer → cross-examination by conducting debate procedure

## [Debate 방식]

1. Agent들로부터 initial response generation
  2. 각 response들은 모두 concatenation
  3. 다시 각 agent들에게 2의 concatenated response를 바탕으로 새로운 response 생성 요청  
→ Refining its own response based on other agents' responses
  4. 1-3 반복
- Consensus in debate: Multiagent debate에서는 convergence가 보장되지 않음  
→ Prompt를 변경해가며 debate duration을 자체적으로 control  
→ 그럼에도 불구하고 불일치되는 경우 voting 방식 채택

# Language Generation through Multiagent Debate

## [Expectation]

- Same answer → confidence increases
- Inconsistent answer → cross-examination by conducting debate procedure

**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:  $\text{Total gems} = 175 + x + 2(x-35) = 3x + 105$ .

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:  $\text{Total gems} = 3(40) + 105 = 225$ . ❌

**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)

**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. ✅

### 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. ✅



# Experiments

- 3 Questions

- (1) Multiagent debate가 reasoning 능력을 어느 정도까지 향상시키는가?
- (2) Multiagent debate가 factual validity를 어느 정도까지 향상시키는가?
- (3) Generation 성능을 향상시키기 위해 어떤 multiagent debate 디자인을 선택하는 게 좋을까?

# Experiments

- Tasks:
  - [Reasoning] Arithmetic, GSM8K, Chess (best next move prediction)
  - [Factuality] Biographies (historical biography generation), MMLU, Chess Move Validity (가능한 다음 move의 타당성 평가)
- Baseline: direct response generation (single), self-reflection, multiple answer generation+voting
- Model: gpt-3.5, (gpt4/llama7b)
- 주로 3개 agents, 2 round로 구성

# Experiments

Model	Arithmetic (%) $\uparrow$	Grade School Math (%) $\uparrow$	Chess ( $\Delta$ 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
Multiagent (Majority)	75.0 $\pm$ 3.9	81.0 $\pm$ 3.9	105.2 $\pm$ 5.9
Multiagent (Debate)	<b>81.8 <math>\pm</math> 2.3</b>	<b>85.0 <math>\pm</math> 3.5</b>	<b>122.9 <math>\pm</math> 7.6</b>

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

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
Multiagent (Majority)	-	67.0 $\pm$ 4.7	36.0 $\pm$ 2.8
Multiagent (Debate)	<b>73.8 <math>\pm</math> 2.3</b>	<b>71.1 <math>\pm</math> 4.6</b>	<b>45.2 <math>\pm</math> 2.9</b>

**Table 2. Multiagent Debate Improves Factual Accuracy.** Multiagent debate improves the factual accuracy.

Model	Arithmetic	GSM8K	MMLU
Single Agent	9.0 $\pm$ 1.6	20.7 $\pm$ 2.3	41.0 $\pm$ 2.8
Single Agent (Reflection)	10.7 $\pm$ 1.7	21.0 $\pm$ 2.3	39.7 $\pm$ 2.8
Multiagent (Majority)	11.0 $\pm$ 1.8	25.7 $\pm$ 2.5	43.3 $\pm$ 2.9
Multiagent (Debate)	<b>13.3 <math>\pm</math> 1.9</b>	<b>29.3 <math>\pm</math> 2.6</b>	<b>47.7 <math>\pm</math> 2.9</b>

**Table A5. Multiagent Debate on chat-Llama 7B.** Our approach also improves the performance of the chat-Llama model.

- Multiagent debate = reflection + multiagent generation  $\rightarrow$  boost in performance

# Experiments

	<b>Question:</b> What is the result of $10+20*23+3-11*18$ ?	<b>Question:</b> 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 ❌
	<b>Question:</b> What is the result of $4+23*6+24-24*12$ ?	<b>Question:</b> 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.

	<b>Question:</b> 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.	<b>Question:</b> 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 7. Illustration of MMLU. Illustration of debate when answering factual tasks. Reasoning omitted.

- Correct answer에 대한 확신을 갖는 장점 + 다른 agent의 reasoning 능력을 바탕으로 정답에 도달

# Analysis

- Number of agents / Rounds of Debate

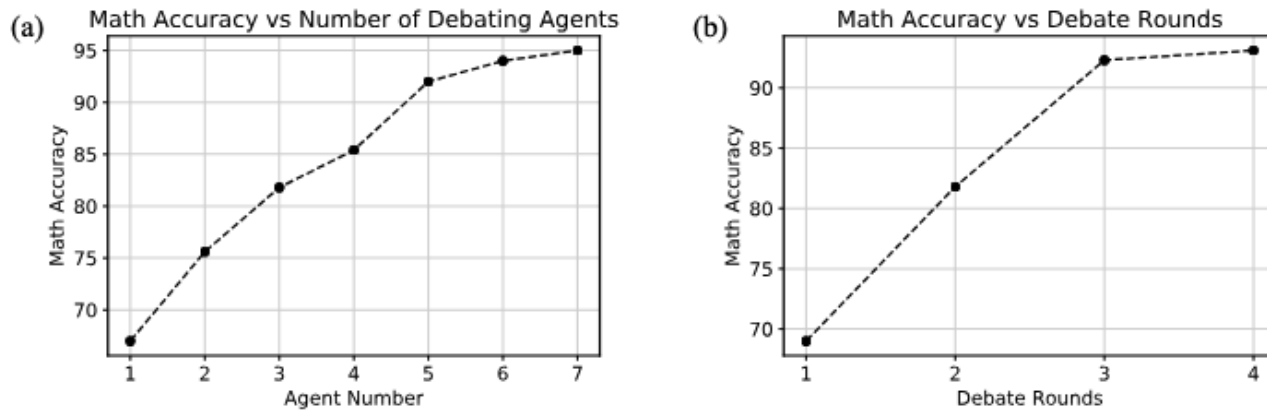


Figure 9. (a) **Performance with Increased Agents.** Performance improves as the number of underlying agents involved in debate increases. (b) **Performance with Increased Rounds.** Performance rises as the number of rounds of underlying debate increases. Analysis in both settings on Arithmetic.

Model	Arithmetic	GSM8K	MMLU
Majority Vote (50 Agents)	92.0 ± 2.7	85.0 ± 3.6	67.0 ± 4.7
Debate (10 Agents 2 Rounds)	<b>96.0 ± 1.3</b>	<b>89.0 ± 3.1</b>	<b>71.0 ± 4.5</b>

Table A4. **Multiagent Debate with Many Agents.** Our approach also improves the performance with a very large number of agents.

- Debate를 위한 agent, round를 늘릴수록 성능에 긍정적 영향
- 심지어는 50 Agents를 활용한 majority voting 방식보다 더 우수한 성능

# Analysis

- **Effect of Debate Length on Accuracy**
- Debate length를 조절할 수 있었음
  - Short debate: 기존 응답에 기반해 update response 생성 (참고 측면)
  - Long debate: 기존 응답들을 사용하여 update response 생성 (활용 측면)
- Longer prompts -> slower convergence to correct answers / better final consensus

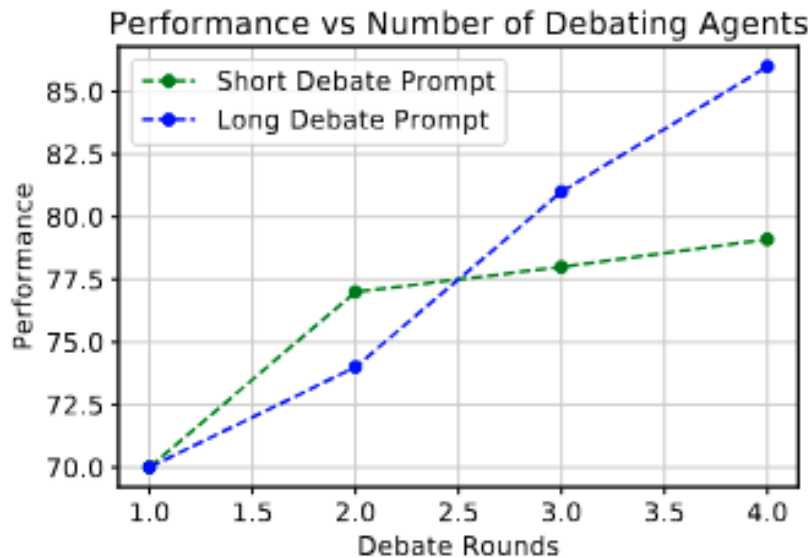


Figure 10. **Performance vs Debate Length.** Prompts which induce longer debate improve performance. Analysis on GSM8K.

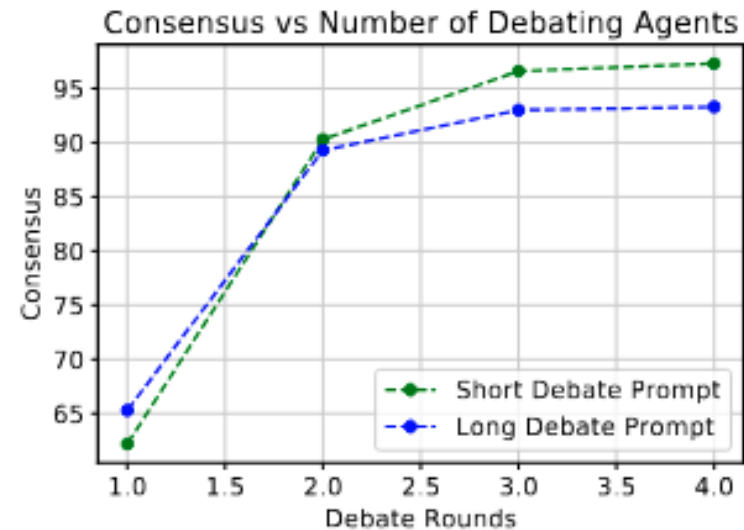
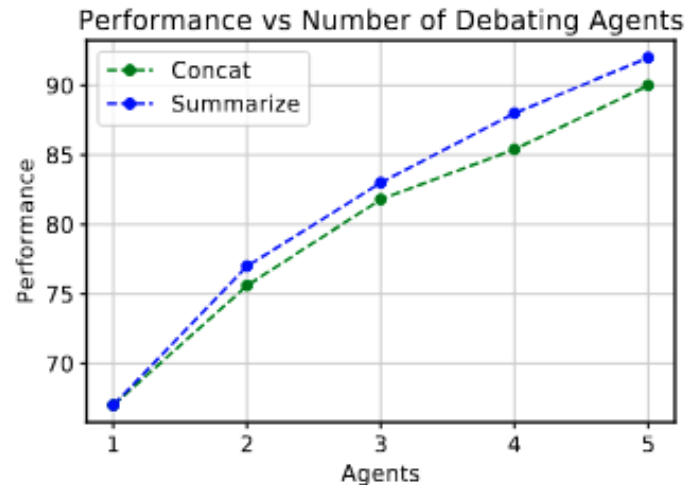


Figure A1. **Effect of Prompts on Consensus.** Using a short debate prompt induces faster consensus between agents. Analysis on GSM8K.

# Analysis

- **Summarization**
- 바로 응답들을 concatenation하는 것 -> expensive / context length가 너무 길어지면 잘림
- 각 response를 요약 후 debate하도록 지시
- 요약 후 debate하는 것이 오히려 성능 향상에 도움을 줌



*Figure 11. 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 on Arithmetic.

# Conclusion

- LLM pretrain은 사실상 불가능. 이제는 기존의 LLM들을 잘 조합하고 활용함으로써 실생활에서 AI의 유용함을 증명해야 하는 단계
- Complexity가 높은 tasks로 구성된 real-world problem에서 앞으로 multiagents를 활용하는 방향으로 흘러갈 가능성이 높음
- Planning 및 task decomposition / subtask 수행 등 과정에서 추가적으로 활용될 debate 방식의 장점을 확인
- Cost 문제는 여전히 해결되어야 함

Method	Arithmetic	GSM	Chess Reasoning	Biography	MMLU	Chess Validity
Single Agent	95.6 ± 5.1	111.5 ± 5.5	8.3 ± 0.6	220.5 ± 3.6	91.7 ± 5.1	39.0 ± 1.1
Single Agent (Reflection)	170.2 ± 5.7	155.2 ± 10.2	64.6 ± 3.2	297.2 ± 11.8	97.2 ± 5.9	92.8 ± 1.6
Multiagent (Majority)	564 ± 10.7	660.1 ± 16.2	49.2 ± 3.0	1295 ± 7.3	422.31 ± 12.3	331.8 ± 2.9
Multiagent (Debate)	548.1 ± 9.4	524.2 ± 11.7	199.5 ± 5.3	967.1 ± 43.7	527.7 ± 17.1	306.1 ± 1.9

Table A9. **Generation Token Cost of Methods on Each Dataset.** Average number of generated tokens (summed across all rounds of debate / conversation) when answering a query per method per dataset.



# 추가 multiagent debate 방법론

- MAD (EMNLP24)

- Affirmative debater
- Negative debater
- Moderator

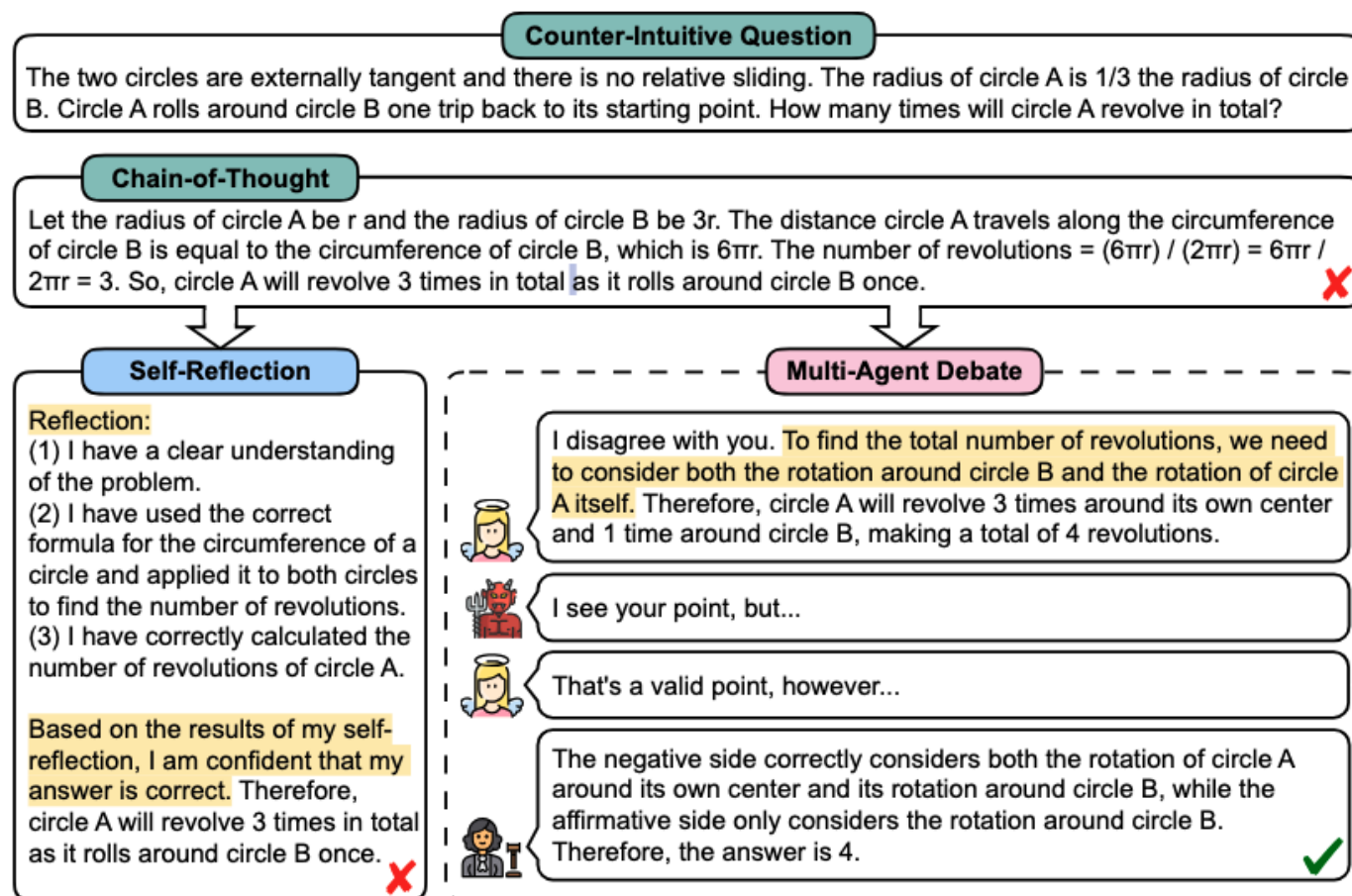


Figure 2: Framework of Multi-Agent Debate. Here we designate the devil (👹) as the affirmative side while the angel (👼) as the negative side. We want the angel to correct the devil's mistakes.

# 추가 multiagent debate 방법론

## • Reconcile (ACL24)

- (1) initial response + explanation + confidence 생성
- (2) Discussion phase
  - (1) results + convincing sample
  - Convincing sample: wrong answer (자신 제외) + human explanation (correction)
    - 다른 agent가 어떤 sample을 틀렸고 왜 잘못되었는지 교정해주는 sample을 추가
    - 각 agent들의 특징을 파악할 수 있게 함
- (3) Team answer generation
  - weighted voting scheme을 활용한 final answer 도출

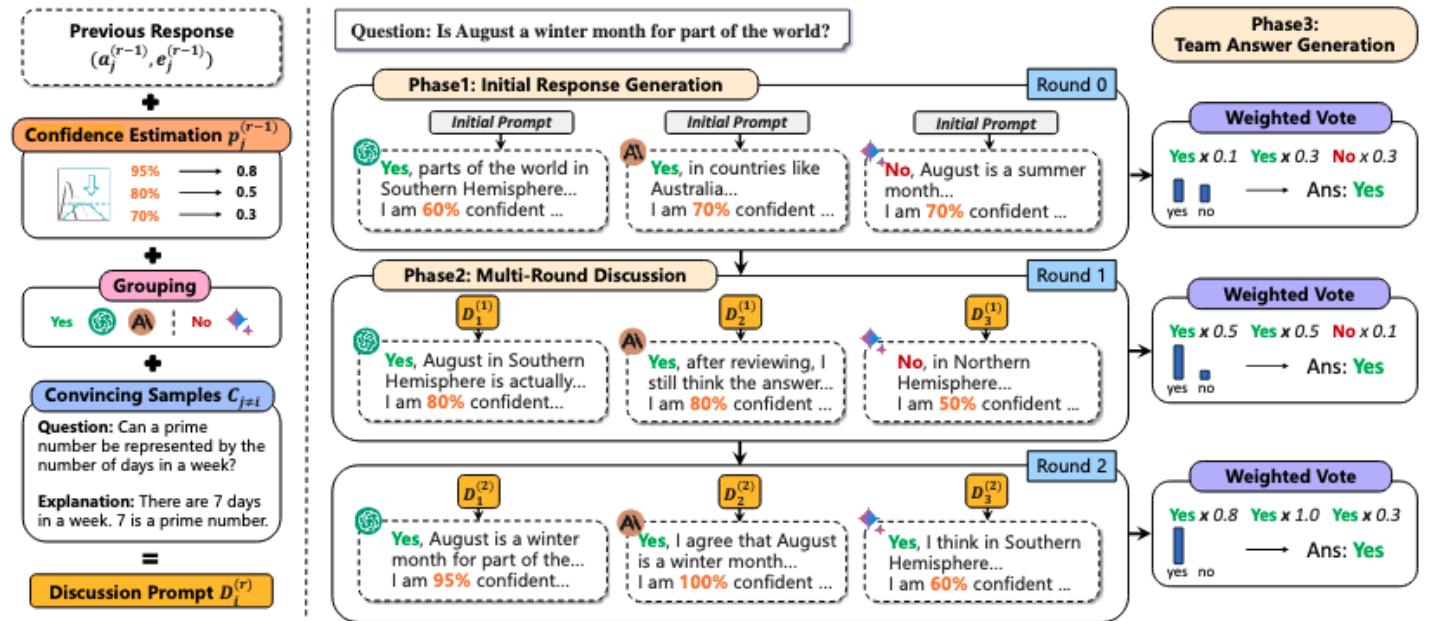


Figure 2: Overview of RECONCILE with ChatGPT, Bard, and Claude2, consisting of three phases: (1) Initial Response Generation: Each agent generates an initial answer and explanation. (2) Multi-Round Discussion: Each model is presented with a discussion prompt (as illustrated on the left) and subsequently generates an updated answer and explanation. (3) Team answer generation: The team answer is determined by a weighted vote at the end of each round. The left part of the figure shows the discussion prompt for an agent, consisting of (a) grouped answers and explanations of all agents from the previous round, (b) estimated confidence, and (c) demonstrations of convincing samples.

# 추가 multiagent debate 방법론

- MoA (Mixture-of-agents) (ICLR25 6/8/8/8)

- Multiple MoA layers

- Proposers
- Aggregator

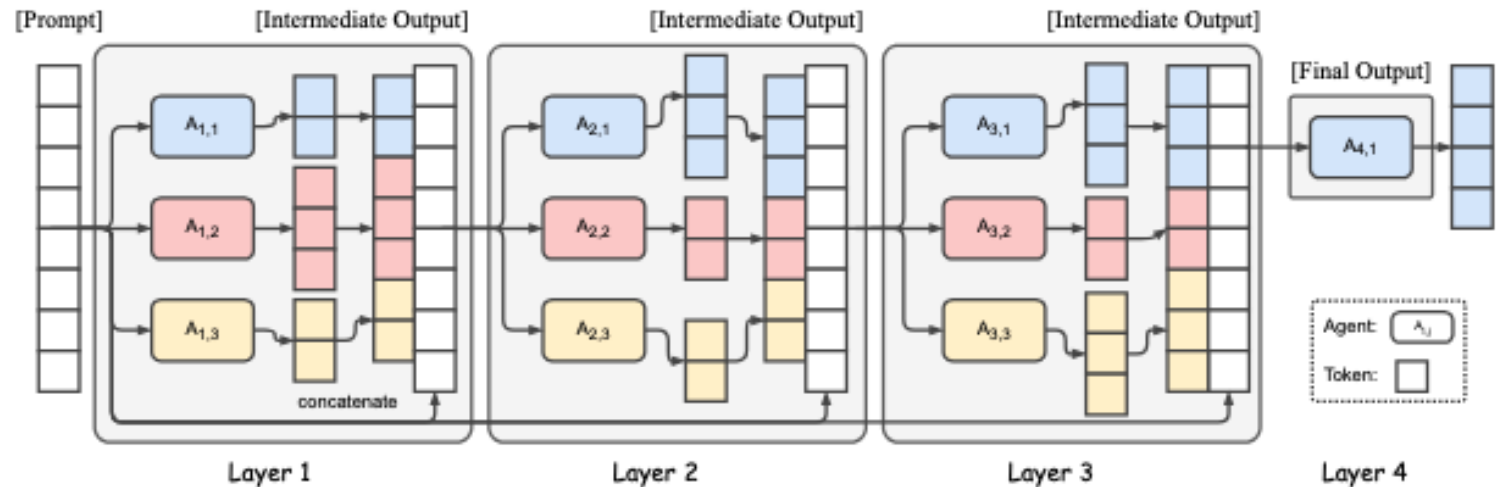


Figure 2: Illustration of the Mixture-of-Agents Structure. This example shows 4 MoA layers where the first layer has 3 proposers, the second and third layer have 3 aggregators that also serve as proposers for the next layer, and the last layer has one aggregator.

# Paper

## **Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?**

**Qineng Wang<sup>1\*</sup>   Zihao Wang<sup>2\*</sup>   Ying Su<sup>2</sup>   Hanghang Tong<sup>3</sup>   Yangqiu Song<sup>2</sup>**

<sup>1</sup>Zhejiang University   <sup>2</sup>HKUST   <sup>3</sup>UIUC

wongqineng73@zju.edu.cn, htong@illinois.edu

{zwanggc, ysuay, yqsong}@cse.ust.hk

# Introduction

- LLM은 여전히 reasoning skill이 부족, hallucination 문제 발생
- Multiagent 방식은 자동화된 방식으로 discussion 진행 → human discussion과 유사 / 능력 향상
- Multiagent 기반 논문들은 single agent보다 우수한 성능을 낸다고 주장 → 과연 모든 면에서 그럴까?
- In-depth analysis 진행
- 새로운 framework인 CMD 제안

# Introduction

- 기존 debate framework

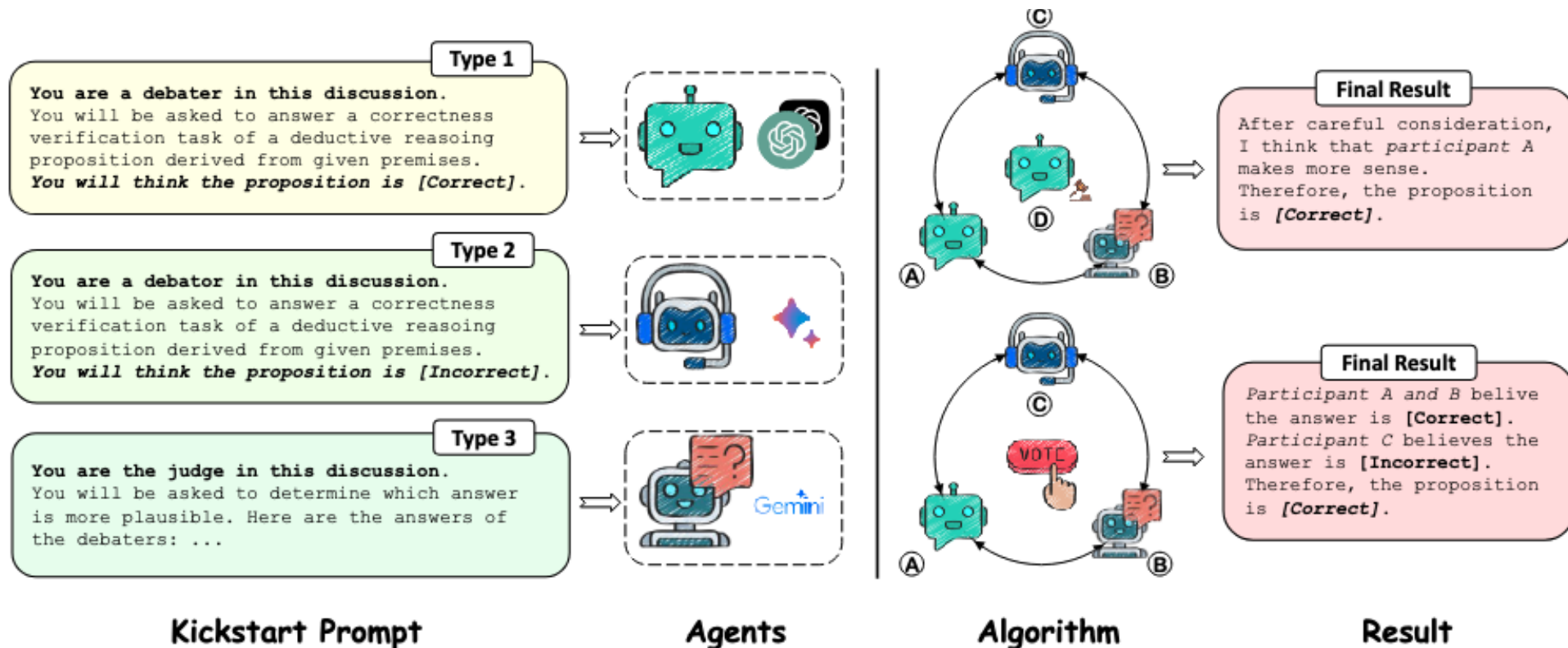


Figure 2: Our proposed design pipeline of multi-agent discussion frameworks. This pipeline operates by having agents starting with a kick-start prompt. Then, agents will start discussion by obeying the rules defined in the algorithm and come to a result in the end.



# CMD: Conquer-and-Merge Discussion

- CMD Framework

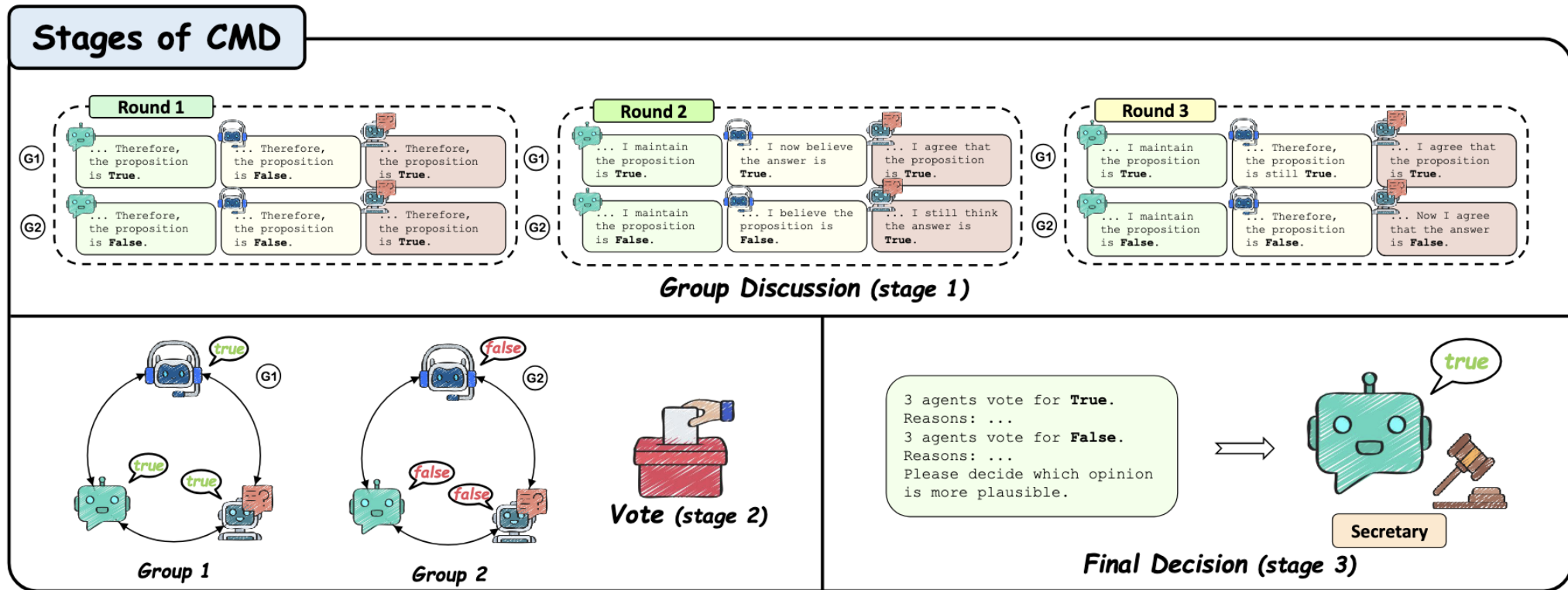


Figure 3: Overview of the Conquer-and-Merge Discussion (CMD) Framework.

# CMD: Conquer-and-Merge Discussion

- CMD Framework
- Three stages: (1) Group discussion (2) Voting (3) Final decision(tie인 경우에만 진행)
- Stage 1: Group discussion
  - Group k 개 생성 – 각각 agent N 개로 구성, 각 agent는 응답 생성
  - 지난 라운드의 각 agent 응답은 다음 라운드에서 참조 가능 (자신 제외)
  - 다른 그룹의 응답을 참고할 수 있지만 explanation이 아닌 정답만 볼 수 있음
- Stage 2 : Voting
  - Maximum round가 넘어가면 모든 agent는 voting을 진행
  - Vote 후 결과가 tie라면 Stag 3로 넘어감
- Stage 3: Final decision
  - Extra agent가 final decision



# Experiments

- Setup
- Models: GPT-3.5, Bard, Gemini Pro
- Maximum discussion round: 3
- Tasks: ECQA(common sense knowledge), GSM8K, FOLIO-wiki(reasoning)

# Analysis of FOLIO-wiki dataset

- Multiagent vs single agent + the strongest prompt

- **Single agent**

- question description, demo

- **Multiagent**

- question description, demo

- demo가 추가되는 경우, 다른 방법론들은 single agent 결과를 이기지 못함

- CMD는 single보다 높은 성능

Prompt Components			Multi-Agent Discussion (%)				Single Agent (%)
Q-Desc.	A-Desc.	Demo.	MAD (3)	Debate (3)	Debate (6)	CMD (6)	
✗	✗	✗	64.13	70.00	69.13	73.26	70.22
✓	✗	✗	74.13	75.65	76.30	74.13	73.26
✗	✓	✗	68.91	71.96	71.74	73.89	71.30
✓	✓	✗	71.96	70.22	70.00	71.09	73.91
✓	✓	✓	74.13	75.65	74.78	77.39	76.09

Table 1: Comparative performance of single-agent settings and multi-agent discussions on FOLIO-wiki using ChatGPT-3.5. Abbreviations are: detailed question descriptions (Q-Desc.), and answer format descriptions (A-Desc.), demonstrations (Demo.). Only the question itself is used as input when prompt components are disabled. The number next to the framework represents the number of agents.

\* Prompt composition (detailed question description)

1. In-depth background of the task
2. Answer format description
3. Task-specific demonstration (meticulous crafted-Q에서 A로 도달할 때까지 필요한 모든 전제를 labeling)

→ well-supported agent can perform on par with discussion frameworks

# Analysis of FOLIO-wiki dataset

- Evaluation on All Tasks

- Demonstration이 추가된 경우 single agent와 multiagent는 비슷한 성능

- Demonstration이 없는 경우 multi가 대부분 우수한 성능

→ Expert knowledge or detailed examples가 없는 (새로운 task) 경우에는 multiagent discussion을 통해 다양한 견해 제공 및 추론 능력 향상 가능

Method	ECQA		GSM8K		FOLIO-wiki		Average	
	Direct	Demo	Direct	Demo	Direct	Demo	Direct	Demo
Single Agent	63.00	67.00	69.00	83.00	70.22	76.09	67.41	75.63
MAD (3 Agents)	55.00	58.00	74.00	78.00	61.25	74.13	63.42	70.04
Debate (3 Agents)	67.00	65.00	78.00	81.00	70.00	75.65	71.67	73.88
Debate (6 Agents)	65.00	64.00	74.00	78.00	69.13	74.78	69.38	72.26
<b>CMD</b> (6 Agents)	64.00	63.00	75.00	83.00	73.26	77.39	70.75	74.46

Table 2: Results for all tasks, with and without demonstration settings included. Using ChatGPT-3.5.

# Two Discussion Error Types: A Case Study

- Multiagent discussion이 잘못된 결론으로 도달하는 결과

## 1. Judge mistake:

- single agent는 옳은 응답 생성, multiagent는 일부가 틀린 응답 생성
- judge의 역할에 따라 오답을 최종 응답으로 취할 가능성
- judge의 역할이 매우 중요해짐 (특히나 tie인 경우)

## 2. Wrong answer propagation:

- 옳은 응답 → 틀린 응답으로 잘못 propagation되는 경우도 빈번히 발생

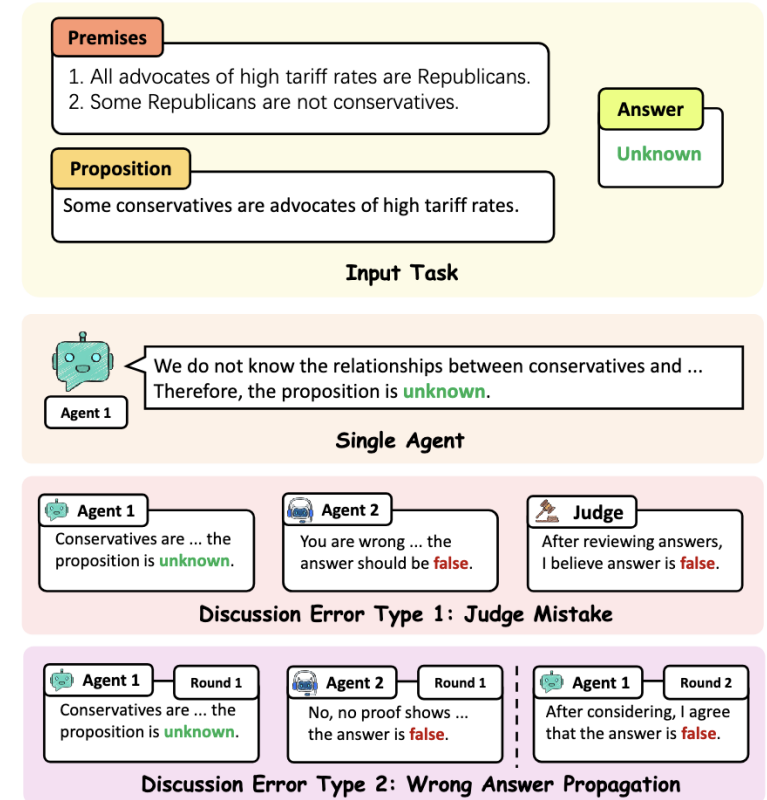


Figure 4: Two common types of errors that may occur in multi-agent discussions are judge mistake and wrong answer propagation. These issues can lead to circumstances where a multi-agent discussion reaches an incorrect conclusion, even if single agent can arrive at the correct one.

# Conclusion

## Takeaways

1. Agent가 충분히 우수한 경우, multiagent discussion이 반드시 reasoning skill을 향상시키는 것은 아님을 밝힘
  - Single agent + strong prompt의 경우 multi-agent discussion과 유사한 성능을 보일 수 있음
2. Demonstration이 없는 경우: multiagent > single agent
3. Two types of common discussion errors:
  - (1) Judge mistake
  - (2) Wrong answer propagation

**Thank you**  
**Q&A**