

AI Agents Debate :Solving Complex Problems with LLMs

DMQA Open Seminar

2025. 11. 28 (금)

Data Mining & Quality Analytics Lab.

박소연

발표자 소개



❖ 박소연 (Soyeon Park)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S Student (2024.09 ~ Present)

❖ Research Interest

- Test-Time Learning
- AI Agent
- Federated Learning

❖ Contact

- syeonpark@korea.ac.kr

Do LLMs work well for all tasks?

❖ LLM은 복잡한 추론 문제 해결에 있어서는 여전히 부족하다.

- LLM은 언어 생성, 요약 같은 표면적 언어 작업에서는 매우 뛰어난 성능을 보임
- 하지만 Complex Reasoning이 필요한 문제 해결에 있어서는 명확한 한계점을 보임



"혜승이는 연필 4자루를 사고, 그 중 2자루를 친구에게 주었다. 그 후 남은 연필의 두배를 더 샀다. 지금 혜승이는 총 몇 자루를 가지고 있을까?"

*정답: 6자루



혜승이는 연필 4자루를 사고
2자루를 주었으니 $4-2=2$, 2자루가 남네.
그런데 "두배를 더 샀다"라고
했으니 $2*2=4$, 총 4자루를 산것이야.

총 4자루의 연필이 있겠다!



처음에 4자루를 사고, 그 중 2자루를
줬으니까 2자루가 남는데...
이후 남은 연필의 두배를 사용했으니까
남은 연필은 $2-4 = -2$.
즉 연필이 없네?

0개의 연필이 있어.



처음에 4자루를 샀는데 2자루를
줬으니까, $4-2=2$ 자루가 남아있고,
남은 연필의 두배를 더 샀으니까
 $2*2=4$ 자루를 추가로 구매한거야.
그러니까 남은 2자루와 새로 산
4자루를 더해야돼.

총 6자루의 연필이 있어!

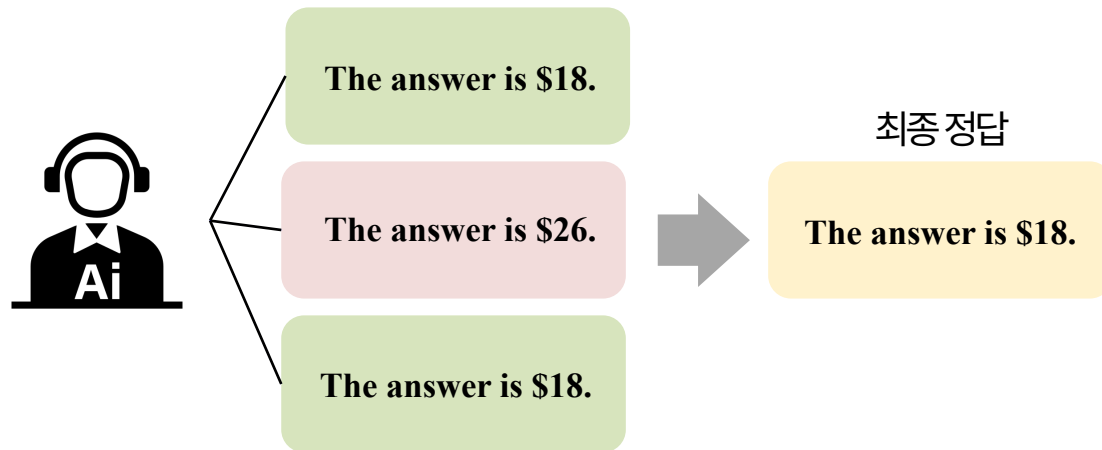
*CoT 적용 답변

Do LLMs work well for all tasks?

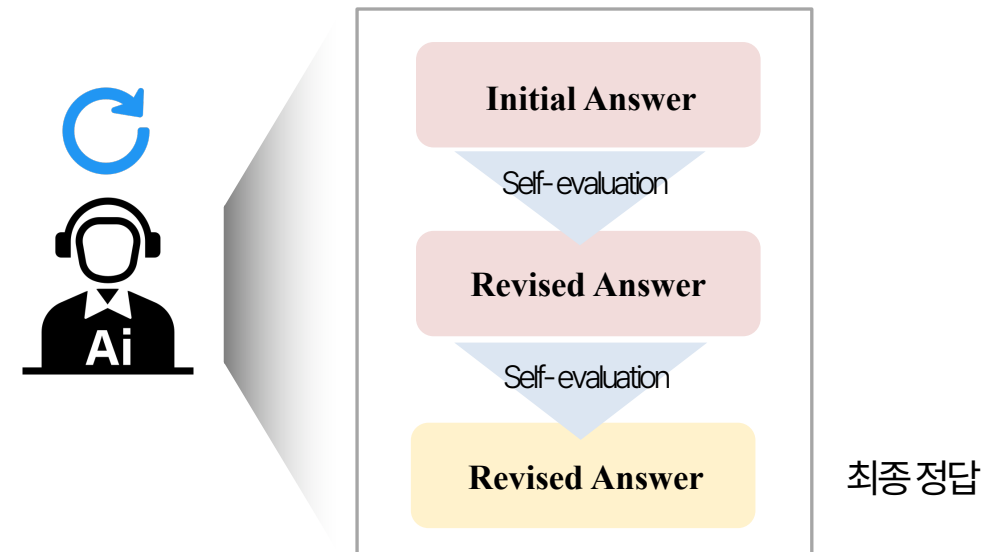
❖ LLM 추론 성능 향상을 위한 기존 접근법

1. **Self-Consistency**: 하나의 LLM에서 여러 개의 답변을 샘플링한 뒤, 다수결 투표 방식을 통해 최종 답변을 결정하는 방식
2. **Self-Reflection**: LLM이 스스로의 답변을 평가하고, 현재 답변이 충분히 만족스럽다고 판단 될 때까지 반복적으로 답변을 개선을 하는 방식

1) Self-Consistency



2) Self-Reflection



Do LLMs work well for all tasks?

❖ LLM 추론 성능 향상을 위한 기존 접근법

1. **Self-Consistency**: 하나의 LLM에서 여러 개의 답변을 샘플링한 뒤, 다수결 투표 방식을 통해 최종 답변을 결정하는 방식
2. **Self-Reflection**: LLM이 스스로의 답변을 평가하고, 현재 답변이 충분히 만족스럽다고 판단 될 때까지 반복적으로 답변을 개선을 하는 방식

1) Self-Consistency



2) Self-Reflection

Why Single LLM is Not Enough for Complex Problems?



Bias Trap

단일 모델의 고정된 사고 패턴에 갇혀
새로운 관점 생성에 한계가 있다

No External Feedback

자기 답에 대한 자기 평가일 뿐
자기 오류를 스스로 판별해내기 어렵다

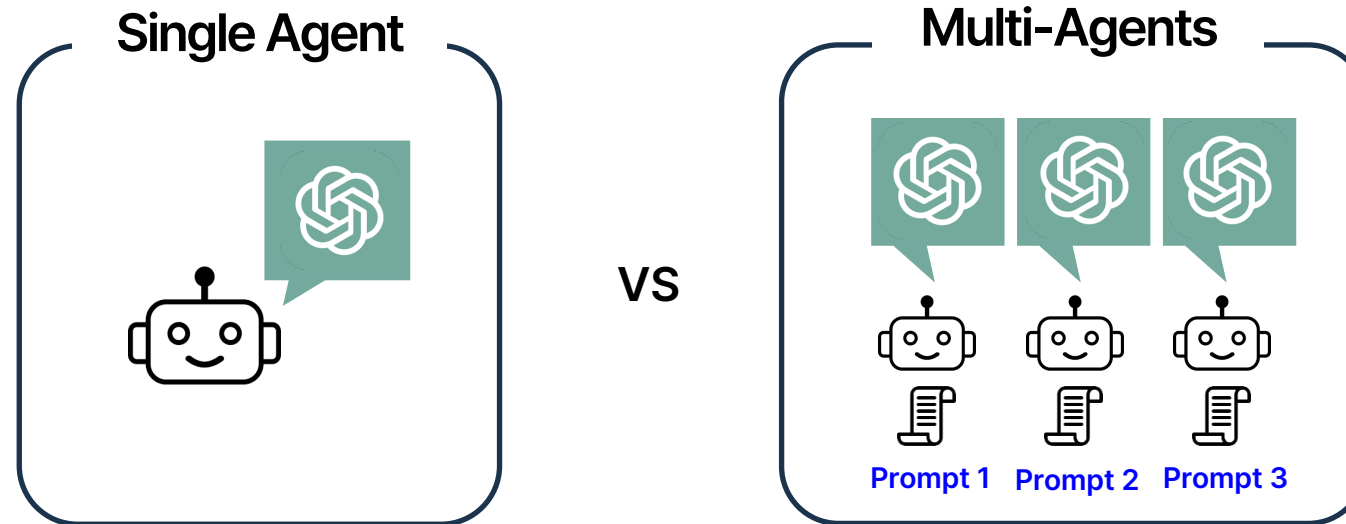
최종 정답

최종 정답

What is "AI Agents Debate"?

❖ AI Agents Debate with Multi-LLMs

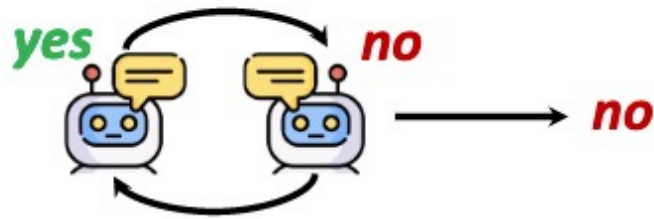
- 복잡한 문제 해결을 위해 여러 LLM 에이전트가 협업/토론/경쟁을 수행하며, 더 나은 결론을 도출하는 방식



How can Multi-Agents be used?

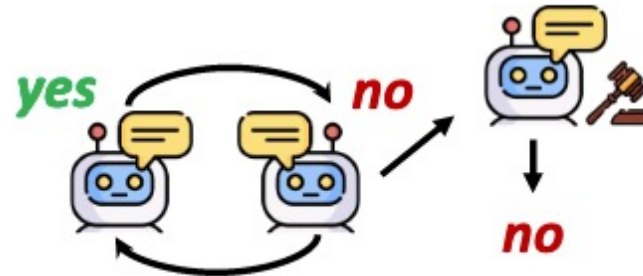
1) Convergence

여러 에이전트가 서로의 답변을 읽고
수정하며 **하나의 결론으로 수렴**



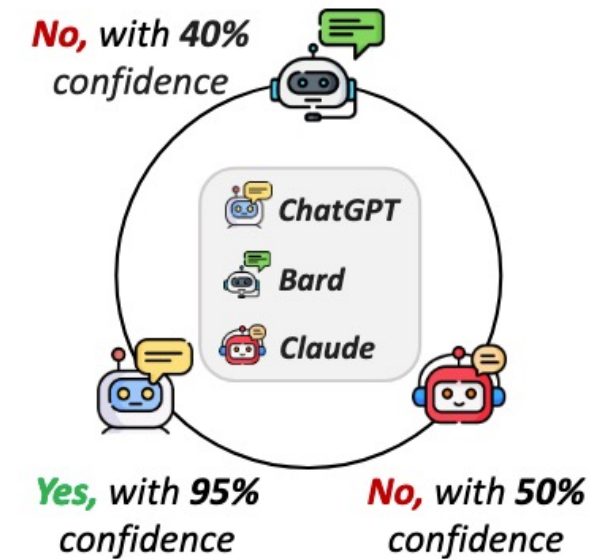
2) Judge

여러 에이전트가 서로 답변을 주고 받은 이후,
별도의 심판 에이전트가 최종 답을 도출



3) Voting

여러 에이전트가 답변을 내고,
다수결이나 가중 투표로 최종 답을 도출



How can Multi-Agents be used?

1) Convergence

여러 에이전트가 서로의 답변을 읽고
수정하며 **하나의 결론으로 수렴**

*Debate

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/.

2) Judge

여러 에이전트가 서로 답변을 주고 받은 이후,
별도의 심판 에이전트가 최종 답을 도출

*MAD

Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate

Tian Liang^{1,3*} Zhiwei He^{2*} Wenxiang Jiao^{3*} Xing Wang^{3†} Yan Wang³
Rui Wang³ Yuju Yang¹ Shuming Shi³ Zhaopeng Tu¹
¹Tsinghua University ²Shanghai Jiao Tong University ³Tencent AI Lab
¹{liangt2@mails, yang_yujiu@sz}, tsinghua.edu.cn
²zwhe.cs@sjtu.edu.cn ³{joelwxjiao, brightxwang, zptu}@tencent.com

Abstract

Modern large language models (LLMs) like ChatGPT have shown remarkable performance on general language tasks but still struggle on complex reasoning tasks, which drives the research on cognitive behaviors of LLMs to explore human-like problem-solving strategies. Along this direction, one representative strategy is self-reflection, which asks an LLM to refine the solution with the feedback generated by itself iteratively. However, our study shows that such reflection-style methods suffer from the *Degeneration-of-Thought (DoT)* problem: once the LLM has established confidence in its solutions, it is unable to generate novel thoughts later through reflection even if its initial stance is incorrect. To address the DoT problem, we propose a **Multi-Agent Debate (MAD)** framework, in which multiple agents express their arguments in the state of "tit for tat" and a judge manages the debate process to obtain a final solution. Clearly, our

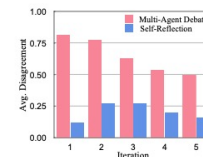


Figure 1: Disagreement between two adjacent iterations with respect to the iteration of debate/self-reflection.
 et al., 2023) but still struggle on complex reasoning tasks (Zhu et al., 2023a; Gou et al., 2023), which drives the research on cognitive behaviors of LLMs to explore human-like problem-solving strategies. In particular, *self-reflection* (Madaan et al., 2024; Shinn et al., 2024), a concept that usually refers to the process of introspection and examination of a person's own thoughts, has been explored to solve

3) Voting

여러 에이전트가 답변을 내고,
다수결이나 가중 투표로 최종 답을 도출

*ReConcile

RECONCILE: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs

Justin Chih-Yao Chen Swarnadeep Saha Mohit Bansal
UNC Chapel Hill
{cychen, swarna, nbansal}@cs.unc.edu

Abstract

Large Language Models (LLMs) still struggle with natural language reasoning tasks. Motivated by the *society of minds* (Minsky, 1988), we propose RECONCILE, a multi-model multi-agent framework designed as a round table conference among diverse LLM agents. RECONCILE enhances collaborative reasoning between LLM agents via multiple rounds of discussion, learning to convince other agents to improve their answers, and employing a confidence-weighted voting mechanism that leads to a better consensus. In each round, RECONCILE initiates discussion between agents via a "discussion prompt" that consists of (a) grouped answers and explanations generated by each agent in the previous round, (b) their confidence scores, and (c) demonstrations of answer-rectifying human explanations, used for convincing other agents. Experiments on seven benchmarks demonstrate that RECONCILE significantly improves LLMs' reasoning – both individually and as a team – surpassing prior single-agent and multi-agent baselines by up to 11.4%, and even outperforming GPT-4 on

self-reflection is an introspective process that allows the model to improve its outputs by generating feedback from the model itself (Madaan et al., 2023; Shinn et al., 2023). However, self-reflection suffers from *Degeneration-of-Thought* – when the model is overly confident in its answer, it is unable to generate novel thoughts even after multiple rounds of feedback (Liang et al., 2023).

To promote more diverse thoughts, past work has drawn inspiration from the concept of *society of minds* in multi-agent systems (Minsky, 1988; Zhuge et al., 2023). It highlights the importance of communication and collaboration between multiple agents for complex decision-making tasks. While such collaborative frameworks like multi-agent debate (Liang et al., 2023; Du et al., 2023) increase the reasoning diversity through the process of a debate, multiple agents have typically been limited to different instances of the same underlying model like ChatGPT (OpenAI, 2022).² This results in an inherent model bias, a restricted knowledge scope, and a lack of external feedback from other

Improving Factuality and Reasoning in Language Models through Multiagent Debate (Google Brain 2023)

DEBATE

❖ Improving Factuality and Reasoning in Language Models through Multiagent Debate

- 2023년 Google Brain, 인용 수 1177회
- 여러 LLM 에이전트가 서로의 답변을 읽고 반박/수정하며 반복적으로 토론한다면, 단일 에이전트보다 우수한 성능을 낼 수 있다!
- 여러 라운드를 거치며 점진적으로 근거와 논리를 강화시키며 최종적으로 하나의 수렴된 답변을 도출하는 구조 제안

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

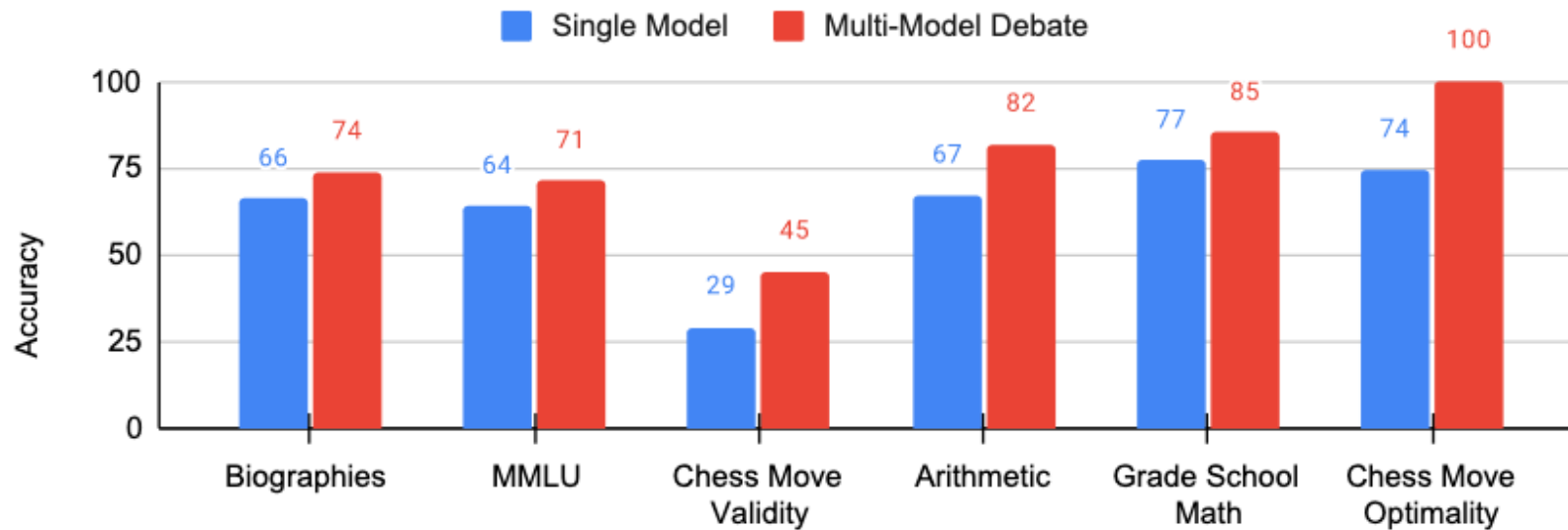
Joshua B. Tenenbaum
MIT CSAIL, BCS, CBMM
jbt@mit.edu

Igor Mordatch
Google Brain
imordatch@google.com

DEBATE

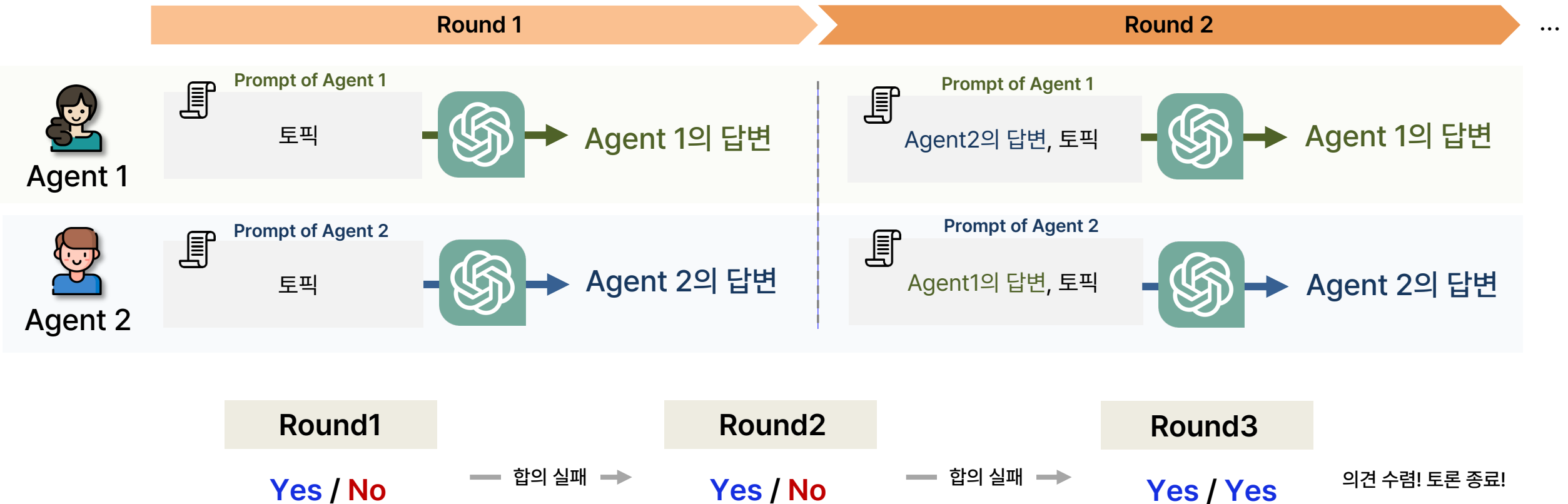
❖ Improving Factuality and Reasoning in Language Models through Multiagent Debate

- 2023년 Google Brain, 인용 수 1177회
- 여러 LLM 에이전트가 서로의 답변을 읽고 반박/수정하며 반복적으로 토론한다면, 단일 에이전트보다 우수한 성능을 낼 수 있다!
- 여러 라운드를 거치며 점진적으로 근거와 논리를 강화시키며 **최종적으로 하나의 수렴된 답변을 도출**



DEBATE

❖ DEBATE 토론 방식: 여러 라운드를 거치며 최종적으로 하나의 수렴된 답변을 도출



DEBATE

❖ DEBATE 토론 방식: 여러 라운드를 거치며 최종적으로 하나의 수렴된 답변을 도출

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.



DEBATE

❖ DEBATE 토론 방식: 여러 라운드를 거치며 최종적으로 하나의 수렴된 답변을 도출

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. ❌

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. ✅

DEBATE

❖ DEBATE 토론 방식: 여러 라운드를 거치며 최종적으로 하나의 수렴된 답변을 도출

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.



의견 수렴 → 토론 종료 !

DEBATE

❖ Experiments (Quantitative Results)

- Arithmetic: 단순 산술 문제 / GSM8K: 수학 추론 문제 / Chess Move Prediction: 체스 다음 수 예측
- Multi-Agent Debate 방식은 3 Agent, 2 Round 만으로도 추론 능력을 크게 향상시키는 결과를 보임

Model	Arithmetic (%) ↑	Grade School Math (%) ↑	Chess (Δ PS) ↑
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

*2 round, 3 debaters

DEBATE

❖ Experiments (Qualitative Results)

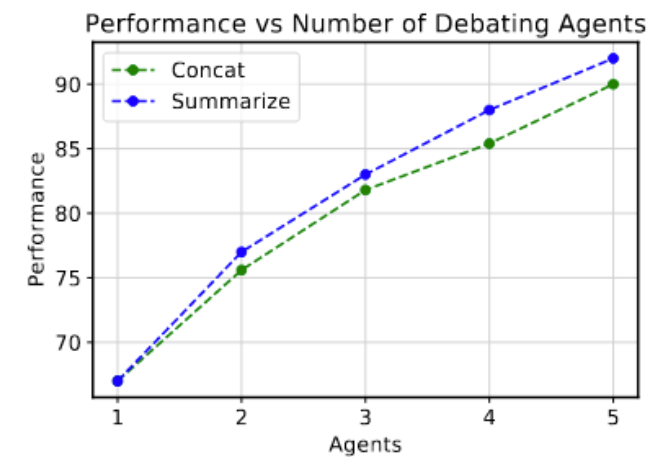
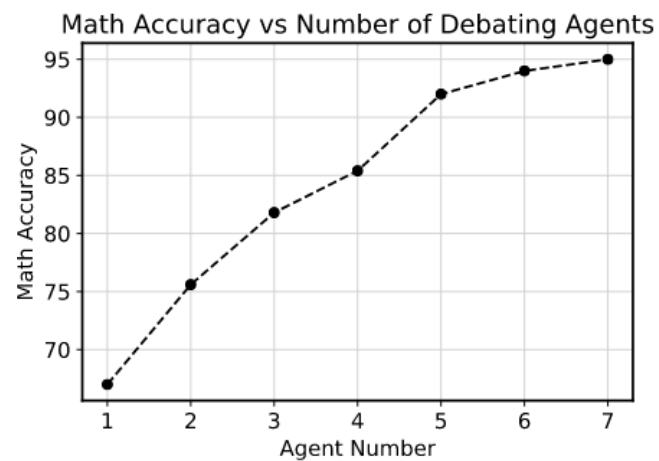
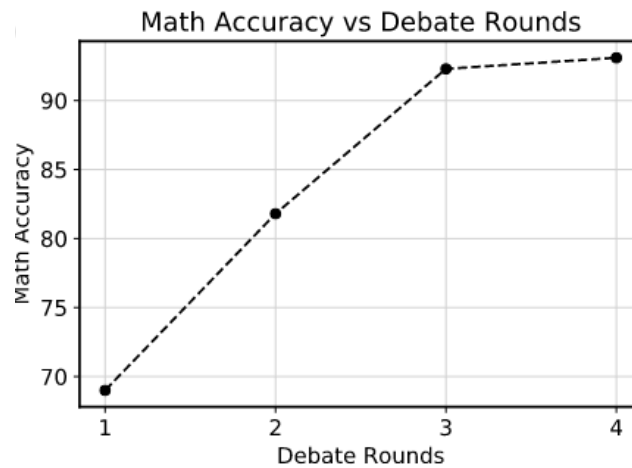
- 모든 모델이 처음에는 잘못된 답을 내더라도, 서로의 추론을 비판하는 과정에서 최종적으로 올바른 답에 도달하는 경우가 존재
- 토론의 목적
 - 올바른 답을 강화하는 것(X)
 - 토론 과정을 통해 정답에 수렴하도록 만드는 것 (O)

	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 ✅

DEBATE

❖ Experiments

- 라운드 수: 증가할수록 성능도 증가하나, 4라운드 초과부터는 거의 동일한 성능
- 에이전트 수: 많을 수록 성능 향상
 - 그러나 참여 에이전트 수가 많아진다면, 다른 에이전트들의 모든 응답을 이어 붙여서 제공하는 방식은 비효율적
 - 따라서, 다른 에이전트들의 응답을 요약하여 제공하는 방식을 사용하면 성능 향상에 도움



Encouraging divergent thinking in large language models through multi-agent debate (EMNLP 2024)

MAD

❖ Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate

- 2024년 EMNLP, 인용 수 776회
- 기존의 Self-Reflection 방식은 스스로 생성한 답변에 대한 자기 확신이 높다면, 이후 라운드에서 더 이상 새로운 생각을 생각해내지 못하는 문제가 있음
- 이를 해결하기 위해, 여러 에이전트가 서로 반박 구조(tit-for-tat)로 토론하고, 별도의 Judge 에이전트가 토론 내역을 보고 최종 답변을 결정하는 구조 제안

Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate

Tian Liang^{1,3*} Zhiwei He^{2*} Wenxiang Jiao^{3*} Xing Wang^{3†} Yan Wang³

Rui Wang² Yujiu Yang^{1†} Shuming Shi³ Zhaopeng Tu³

¹Tsinghua University ²Shanghai Jiao Tong University ³Tencent AI Lab

¹{liangt21@mails, yang.yujiu@sz}.tsinghua.edu.cn

²zwhe.cs@sjtu.edu.cn ³{joelwxjiao, brightxwang, zptu}@tencent.com

MAD

❖ Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate

- 2024년 EMNLP, 인용 수 776회
- 기존의 Self-Reflection 방식은 스스로 생성한 답변에 대한 자기 확신이 높다면, 이후 라운드에서 더 이상 새로운 생각을 생각해내지 못하는 문제가 있음
- 이를 해결하기 위해, 여러 에이전트가 서로 반박 구조(tit-for-tat)로 토론하고, 별도의 Judge 에이전트가 토론 내역을 보고 최종 답변을 결정하는 구조 제안



*앞라운드와 서로 의견이 다르면 1, 동일하면 0
= 즉, 1에 가까울수록 다양한 사고방식을 가짐을 의미

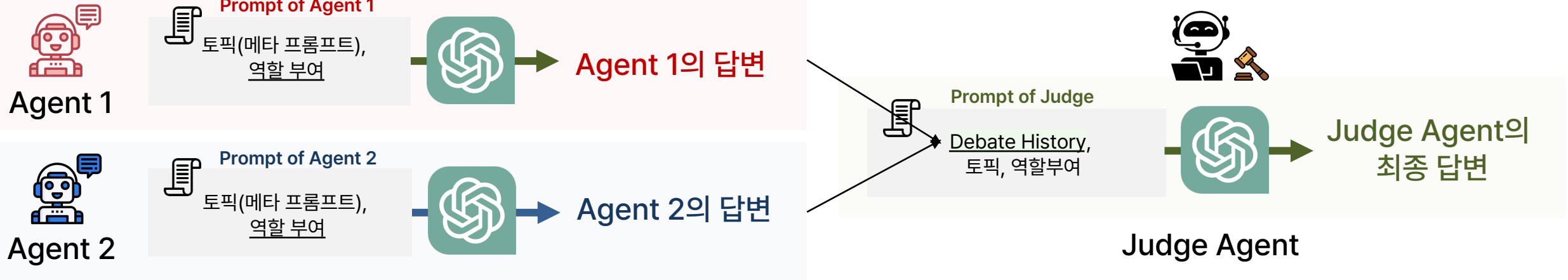
Once the LLM-based agent has established confidence in its answers, it is unable to generate novel thoughts later through self-reflection even if the initial stance is incorrect.

MAD

❖ **MAD 토론 방식:** 여러 에이전트가 서로 반박 구조(tit-for-tat)로 토론하고, 별도의 Judge 에이전트가 토론 내역을 보고 최종 답변을 결정

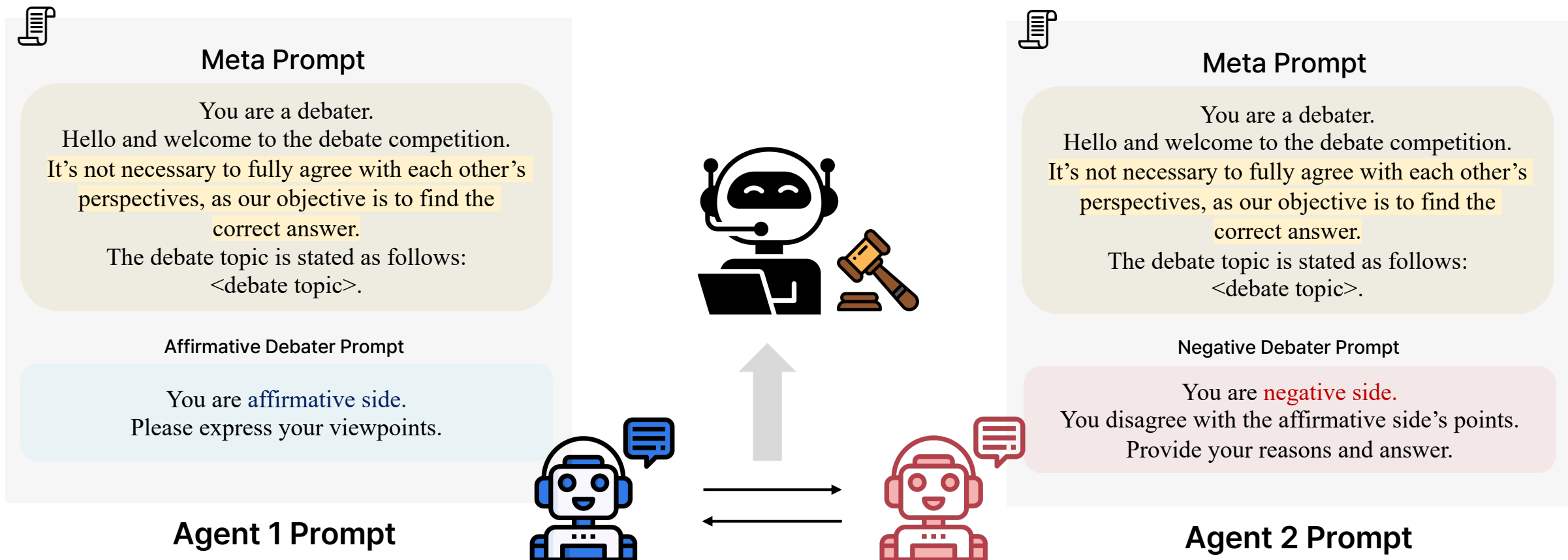
Round 1

...



MAD

❖ **MAD 토론 방식:** 여러 에이전트가 서로 반박 구조(tit-for-tat)로 토론하고, 별도의 Judge 에이전트가 토론 내역을 보고 최종 답변을 결정



MAD

❖ **MAD 토론 방식:** 여러 에이전트가 서로 반박 구조(tit-for-tat)로 토론하고, 별도의 Judge 에이전트가 토론 내역을 보고 최종 답변을 결정

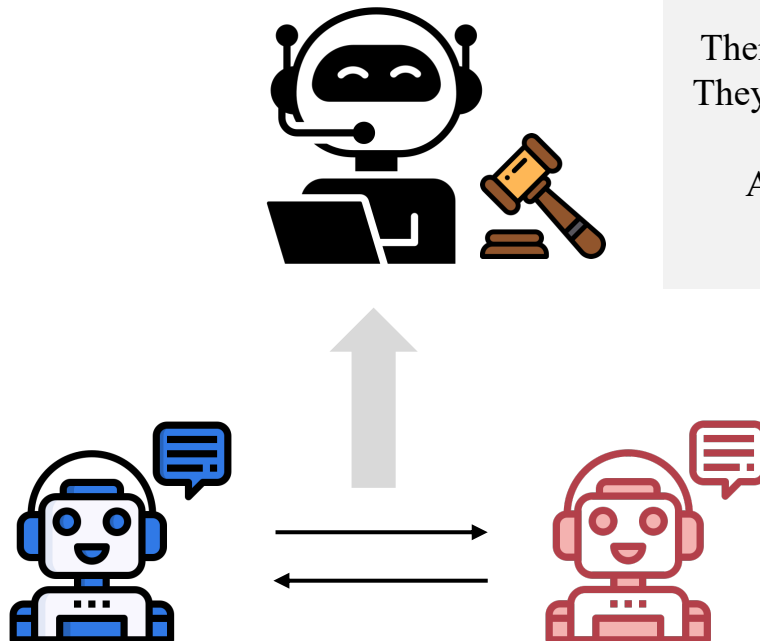
Judge Prompt



You are a **moderator**.

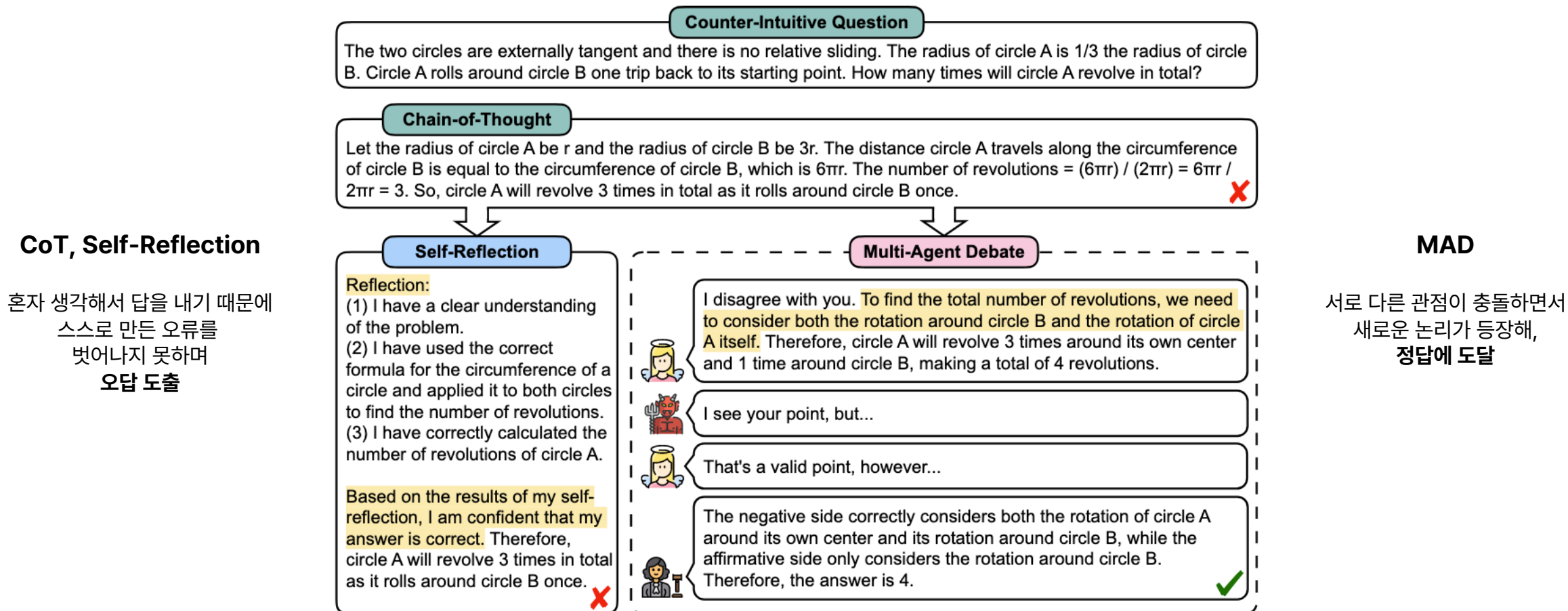
There will be two debaters involved in a debate competition. They will present their answers and discuss their perspectives on the <debate topic>.

At the end of each round, you will evaluate both sides' answers and decide which one is correct.



MAD

❖ **MAD 토론 방식:** 여러 에이전트가 서로 반박 구조(tit-for-tat)로 토론하고, 별도의 Judge 에이전트가 토론 내역을 보고 최종 답변을 결정



MAD

❖ Experiments (commonsense machine translation)

- 중국어 → 영어 번역 데이터셋 (올바른 번역을 위해서는 직역이 아니라 이해가 필수적인 데이터셋)
- GPT-3.5-Turbo + MAD 적용 시, 더 강력한 모델인 GPT-4보다도 성능을 달성함
- 오픈 소스 모델에서도 MAD 적용 시, 성능 향상

* 2 debaters, 1 judge, 3 round

Method	Lexical			Contextless			Contextual		
	COMET	BLEURT	HUMAN	COMET	BLEURT	HUMAN	COMET	BLEURT	HUMAN
GPT-4	82.0	70.1	3.41	84.7	73.6	3.63	85.0	73.7	3.65
Turbo	80.3	68.2	3.14	84.0	72.9	3.43	84.9	73.4	3.57
+ Rerank	80.9	68.6	3.16	84.5	73.2	3.46	85.3	73.9	3.58
+ MAPS	81.9	70.1	3.43	84.2	73.5	3.45	85.2	74.0	3.56
+ Self-Reflect	81.0	69.1	3.43	83.6	72.2	3.46	84.9	73.5	3.63
+ MAD	82.0	70.9	3.78	84.8	73.7	3.67	85.3	74.0	3.67
Vicuna-7b	74.9	62.0	2.55	78.3	64.6	2.53	80.2	68.2	3.23
+ MAD	75.6	62.6	2.67	78.6	66.0	2.69	81.8	69.9	3.27
Vicuna-13b	76.6	63.7	2.81	77.6	66.8	3.04	82.2	70.0	3.37
+ MAD	77.2	65.1	2.96	80.1	67.3	3.11	82.6	70.9	3.45

MAD

❖ Experiments (commonsense machine translation)

- 중국어 → 영어 번역 데이터셋 (올바른 번역을 위해서는 직역이 아니라 이해가 필수적인 데이터셋)
- GPT-3.5-Turbo, GPT-4 모두 단어를 잘못 직역하여 번역
- 그러나, GPT-3.5-Turbo + MAD 적용 시, 다양한 사고 경로를 탐색하기 때문에, 제대로 된 의역을 생성하여 올바르게 번역

Source	吃掉敌人一个师。
Correct Ref.	Destroy a division of the enemy.
Incorrect Ref.	Eat up an enemy division.
GPT-4	Eat up an enemy division.
GPT-3.5-Turbo	Eat up an enemy division.
+ Self-Reflect	Eat up an enemy division.
+ MAD	Eliminate an enemy division.

MAD

❖ Experiments (Counter-Intuitive Arithmetic Reasoning)

- 기존 산술 추론 데이터셋(GSM8K, MultiArith) 보다 훨씬 어려운 데이터셋
 - 잘못된 직관을 억제하고, 다단계 추론이 필요한 데이터셋
- GPT-3.5-Turbo + MAD 적용 시, GPT-4 수준에는 미치지 못하였으나, 다른 모든 비교 기법들을 능가하는 성능을 보임

Method	ACC (%)
GPT-4	51.0
GPT-3.5-Turbo	26.0
+ CoT	28.0
+ Self-Consistency	29.5
+ Self-Reflect	27.5
+ MAD	37.0

MAD

❖ Analysis of Judge

- Strong Debaters + Weak Judge의 성능 > Weak Debaters + Strong Judge의 성능
- Judge는 자신과 동일한 LLM을 사용한 Debater의 답변을 더 자주 선택하는 경향이 있음
 - 만약 모든 에이전트의 LLM이 동일하다면, Negative side를 더 많이 선택하는 경향이 있음

Judge LLM	COMET	HUMAN
Vicuna-13b as Debaters		
Vicuna-13b	79.9	3.20
GPT-3.5-Turbo	80.4	3.25
GPT-3.5-Turbo as Debaters		
Vicuna-13b	83.2	3.47
GPT-3.5-Turbo	84.4	3.69

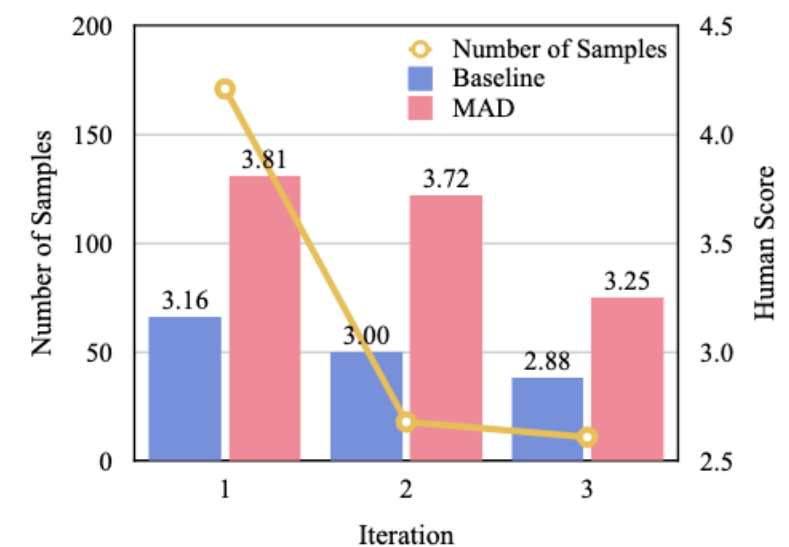
ID	Jud	Debater		Winner		
		Aff	Neg	Aff	Neg	Tie
①	Turbo	Turbo	Turbo	87	104	9
②	GPT-4	GPT-4	GPT-4	67	124	9
③	GPT-4	Turbo	GPT-4	52	136	12
④		GPT-4	Turbo	120	77	3

MAD

❖ Analysis of Debaters

- Debater의 수 ↑ → 장문 처리 능력이 약한 모델에서는 성능 저하
 - 토론자 수가 증가하면 토론 내용의 길이와 복잡성이 크게 증가하여, Judge는 더 많은 양의 텍스트에서 정보를 요약하여 받아들이기 어렵게 됨
- 적절한 수준의 Debater 간 맞대응(tit for tat)은 성능 향상에 필수적
- 복잡한 질문일수록 더 많은 토론(Round)가 필요

# of Debaters	COMET	HUMAN
2 (Default)	84.4	3.69
3	83.1	3.58
4	82.9	3.49



**ReConcile:
Round-table conference improves reasoning
via consensus among diverse LLMs
(ACL 2024)**

ReConcile

❖ RECONCILE: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs

- 2024년 ACL, 인용 수 207회
- Multi-Agent Debate라고 하더라도, 실제로는 대부분 동일 종류의 모델에 기반하여 토론 진행 → 다른 계열의 모델로부터 얻을 수 있는 피드백 부재
- 서로 다른 종류의 LLM 모델을 활용하여 생성한 각 답변에 대하여 신뢰도 기반 가중 투표를 통해 최종 답을 도출하는 구조

RECONCILE: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs

Justin Chih-Yao Chen Swarnadeep Saha Mohit Bansal

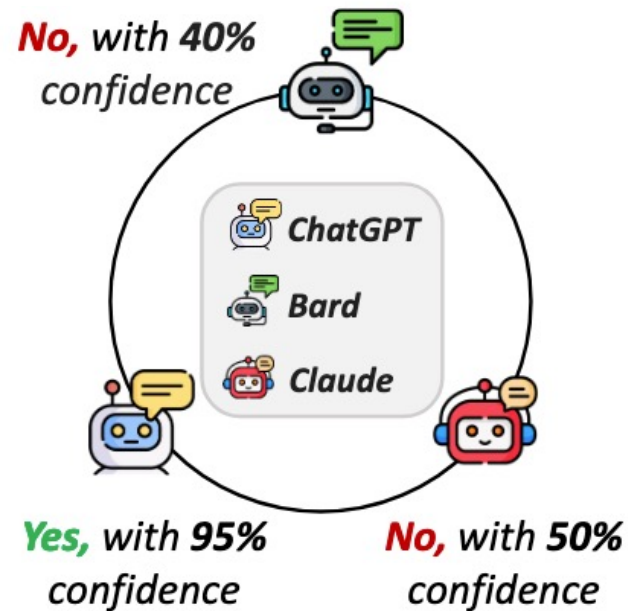
UNC Chapel Hill

{cychen, swarna, mbansal}@cs.unc.edu

ReConcile

❖ RECONCILE: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs

- 2024년 ACL, 인용 수 207회
- Multi-Agent Debate라고 하더라도, 실제로는 대부분 동일 종류의 모델에 기반하여 토론 진행 → 다른 계열의 모델로부터 얻을 수 있는 피드백 부재
- 서로 다른 종류의 LLM 모델을 활용하여 생성한 각 답변에 대하여 신뢰도 기반 가중 투표를 통해 최종 답을 도출하는 구조



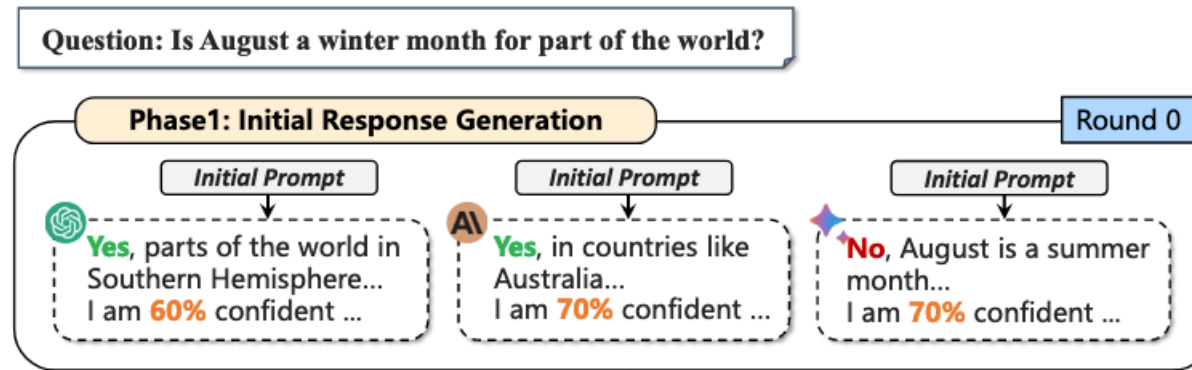
ReConcile

❖ RECONCILE 토론 방식



Initial Prompt

질문, Confidence scoring



ReConcile

❖ RECONCILE 토론 방식

Discussion Prompt

Initial Prompt

질문, Confidence scoring

이전 라운드 토론 히스토리
(각 에이전트의 답변/근거,
Confidence Score)

Question: Is August a winter month for part of the world?

Phase1: Initial Response Generation

Round 0

Initial Prompt



Yes, parts of the world in Southern Hemisphere...
I am **60%** confident ...



Yes, in countries like Australia...
I am **70%** confident ...



No, August is a summer month...
I am **70%** confident ...

Phase2: Multi-Round Discussion

Round 1

$D_1^{(1)}$



Yes, August in Southern Hemisphere is actually...
I am **80%** confident...



Yes, after reviewing, I still think the answer...
I am **80%** confident ...



No, in Northern Hemisphere...
I am **50%** confident ...

ReConcile

❖ RECONCILE 토론 방식

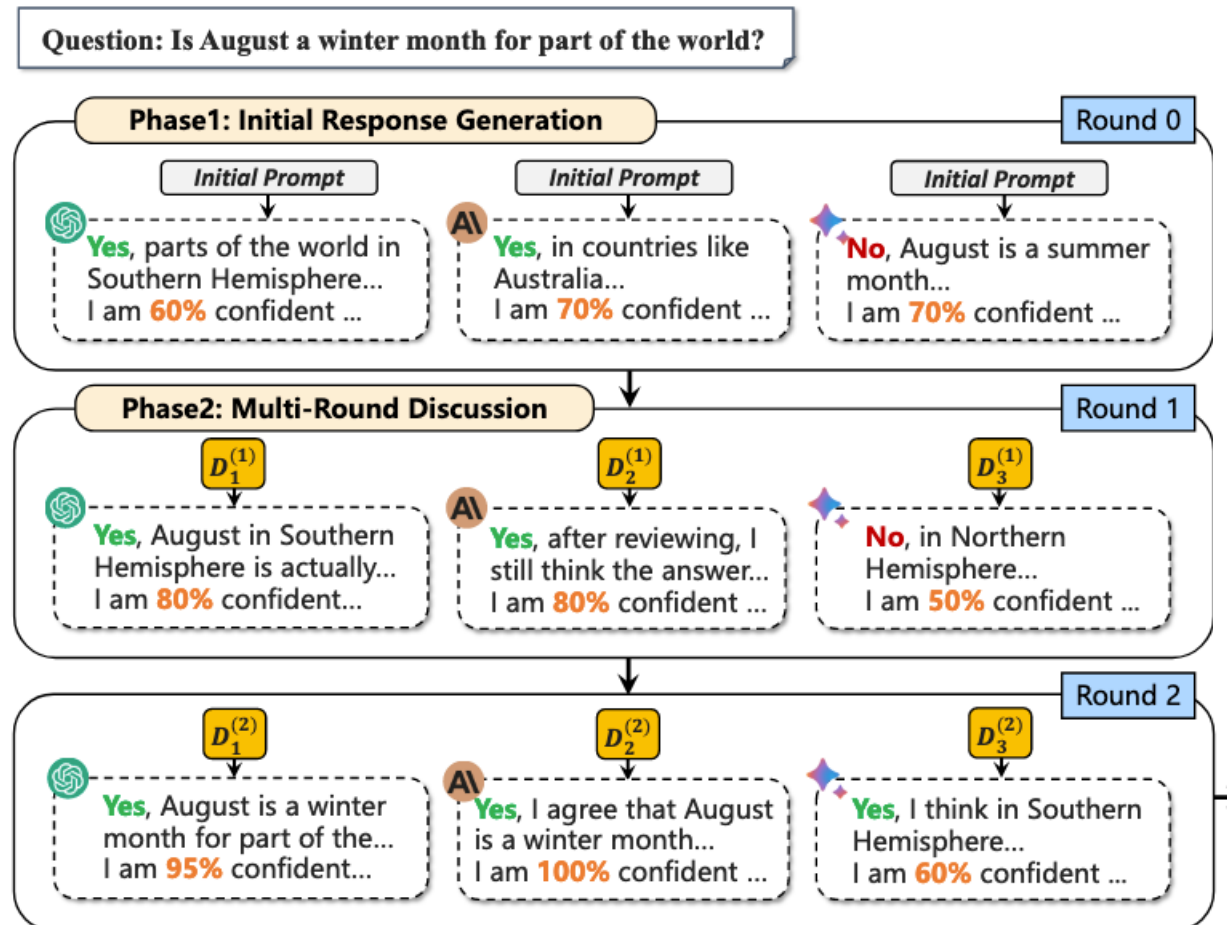
Discussion Prompt

Initial Prompt

질문, Confidence scoring

이전 라운드 토론 히스토리
(각 에이전트의 답변/근거,
Confidence Score)

$$f(p_i^{(r)}) = \begin{cases} 1.0, & \text{if } p_i^{(r)} = 1.0 \\ 0.8, & \text{if } 0.9 \leq p_i^{(r)} < 1.0 \\ 0.5, & \text{if } 0.8 \leq p_i^{(r)} < 0.9 \\ 0.3, & \text{if } 0.6 < p_i^{(r)} < 0.8 \\ 0.1, & \text{otherwise} \end{cases}$$



ReConcile

❖ RECONCILE 토론 방식

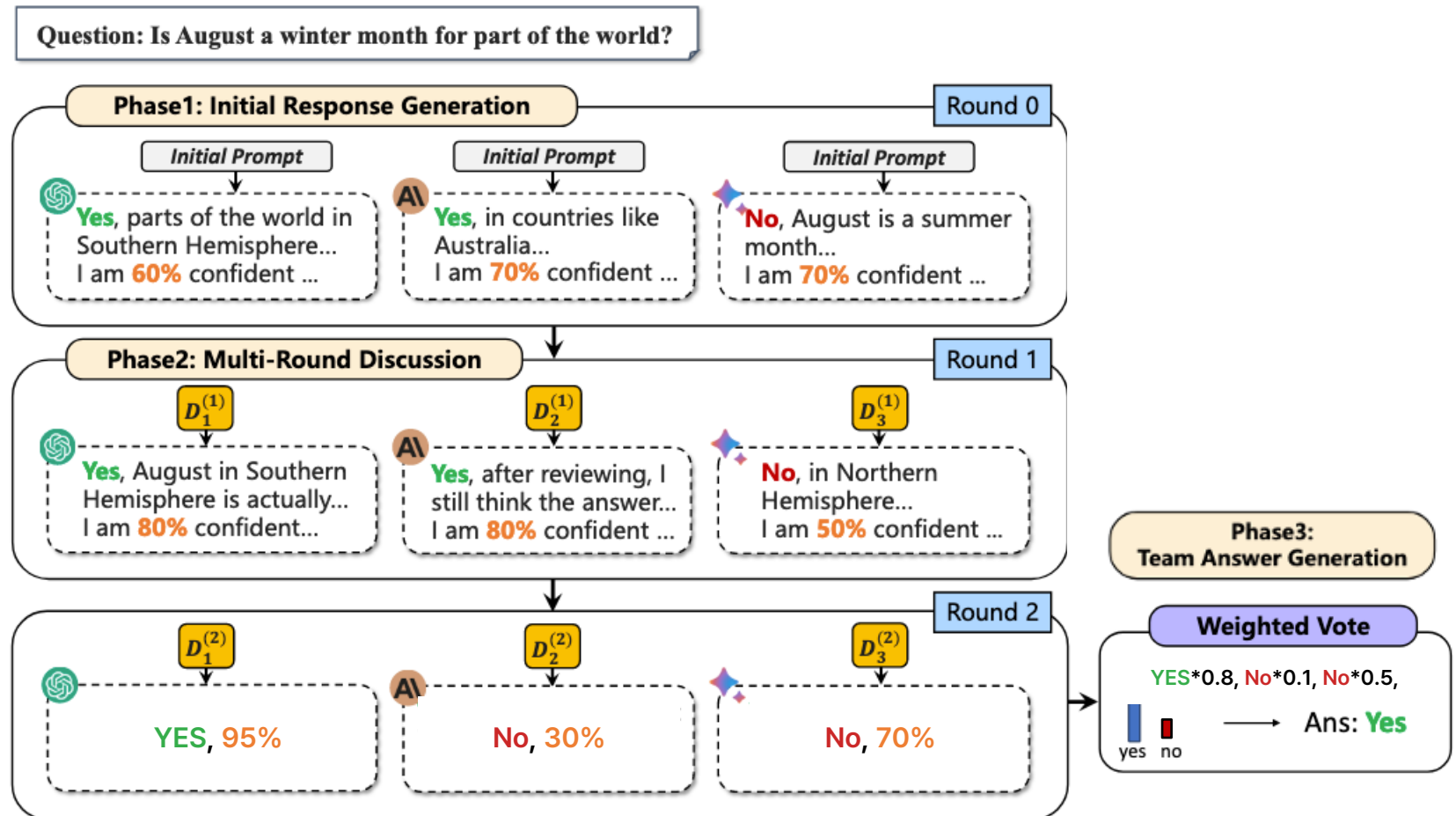
Discussion Prompt

Initial Prompt

질문, Confidence scoring

이전 라운드 토론 히스토리
(각 에이전트의 답변/근거,
Confidence Score)

$$f(p_i^{(r)}) = \begin{cases} 1.0, & \text{if } p_i^{(r)} = 1.0 \\ 0.8, & \text{if } 0.9 \leq p_i^{(r)} < 1.0 \\ 0.5, & \text{if } 0.8 \leq p_i^{(r)} < 0.9 \\ 0.3, & \text{if } 0.6 < p_i^{(r)} < 0.8 \\ 0.1, & \text{otherwise} \end{cases}$$



ReConcile

❖ Experiments

- **Commonsense**(StrategyQA, CSQA) / **Math**(GSM8K, AQuA, MATH) / **Logical Reasoning** (Date)
- 다섯 개의 모든 데이터셋에서 ReConcile은 동일 모델 기반의 단일 에이전트/ 다중 에이전트 베이스라인을 모두 능가
- GPT-4를 사용하지 않았음에도 Commonsense task에서는 GPT-4보다 높은 성능을 달성

Method Category	Method	Agent	StrategyQA	CSQA	GSM8K	AQuA	Date
Vanilla Single-agent	Zero-shot CoT	GPT-4	75.6±4.7	73.3±0.4	90.7±1.7	65.7±4.6	89.0±2.2
	Zero-shot CoT	ChatGPT	67.3±3.6	66.0±1.8	73.7±3.1	44.7±0.5	67.7±1.2
	Zero-shot CoT	Bard	69.3±4.4	56.8±2.7	58.7±2.6	33.7±1.2	50.2±2.2
	Zero-shot CoT	Claude2	73.7±3.1	66.7±2.1	79.3±3.6	60.3±1.2	78.7±2.1
	Eight-shot CoT	Claude2	74.3±0.8	68.3±1.7	84.7±0.9	64.7±1.2	78.7±1.7
Advanced Single-agent	Self-Refine (SR)	ChatGPT	66.7±2.7	68.1±1.8	74.3±2.5	45.3±2.2	66.3±2.1
	Self-Consistency (SC)	ChatGPT	73.3±0.5	73.0±0.8	82.7±0.5	60.3±1.2	69.3±0.4
	SR + SC	ChatGPT	72.2±1.9	71.9±2.1	81.3±1.7	58.3±3.7	68.7±1.2
Single-model Multi-agent	Debate	×3	66.7±3.1	62.7±1.2	83.0±2.2	65.3±3.1	68.0±1.6
	Debate	×3	65.3±2.5	66.3±2.1	56.3±1.2	29.3±4.2	46.0±2.2
	Debate	×3	71.3±2.2	68.3±1.7	70.7±4.8	62.7±2.6	75.3±3.3
	Debate+Judge	×3	69.7±2.1	63.7±2.5	74.3±2.9	57.3±2.1	67.7±0.5
Multi-model Multi-agent	RECONCILE	, ,	79.0±1.6	74.7±0.4	85.3±2.2	66.0±0.8	86.7±1.2

*3 round, temperature 1(단일 모델 MAD)

ReConcile

❖ Experiments

1. ReConcile은 다양한 성능 수준의 에이전트에도 잘 일반화된다
2. ReConcile은 도메인 특화 에이전트에도 잘 일반화된다

*StrategyQA Dataset

Method	Accuracy	
Best Single-agent (zero-shot)	75.6 ()	73.7 ()
Best Multi-agent (Debate)	83.7 ( × 3)	71.3 ( × 3)
RECONCILE	87.7 ( ,  , )	78.0 ( ,  , )

성능이 더 강한 에이전트도
비교적 약한 에이전트로부터
유용한 외부 피드백을 얻어 성능 향상 가능!

*MATH Dataset

Method	Accuracy
GPT-4 (zero-shot)	44.0 ()
Best Single-agent (zero-shot)	50.5 ()
Best Multi-agent (Debate)	48.7 ( × 3)
RECONCILE	58.3 ( ,  , )

도메인 특화 에이전트 또한
도메인에 특화되지 않은 에이전트로부터
유용한 외부 피드백을 얻어 성능 향상 가능!

ReConcile

❖ Experiments (답변 간 유사도 분석)

- 서로 다른 종류의 모델이 Debater로 참여했을 때의 설명은 같은 모델이 생성한 설명보다 더 다양하다

Metric	Method	Accuracy	D (A1, A2)	D (A1, A3)	D (A2, A3)	D (A1, A2, A3)
BERTScore	RECONCILE (🌀 Paraphrased)	72.2	0.9364	0.9376	0.9453	0.9398
	RECONCILE (🌀 ×3)	72.2	0.9077	0.9181	0.9049	0.9102
	RECONCILE (🌀, 🌟, AI)	79.0	0.8891	0.8833	0.8493	0.8739

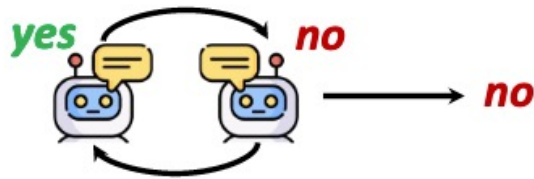
Conclusion

❖ What is Multi-Agent Debate?

: 복잡한 문제 해결을 위해 여러 LLM 에이전트가 협업/토론/경쟁을 수행하며, 더 나은 결론을 도출하는 방식

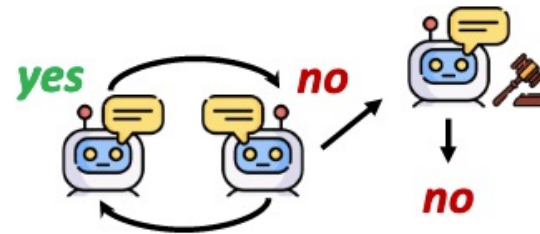
1) DEBATE

여러 에이전트가 서로의 답변을 읽고
수정하며 **하나의 결론으로 수렴**



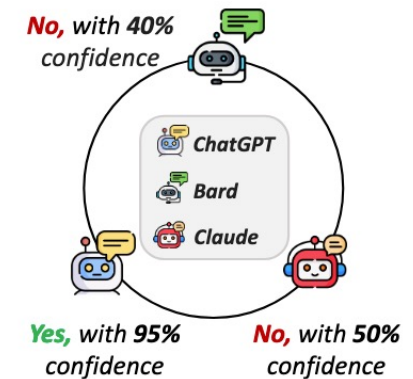
2) MAD

여러 에이전트가 서로 답변을 주고 받은 이후,
별도의 심판 에이전트가 최종 답을 도출



3) ReConcile

서로 다른 종류 모델을 활용한
에이전트가 답변을 내고,
가중 투표로 최종 답을 도출



Thank you

