CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent
Perception and Computing

模式识别国家重点实验室
National Lab of
Pattern Recognition

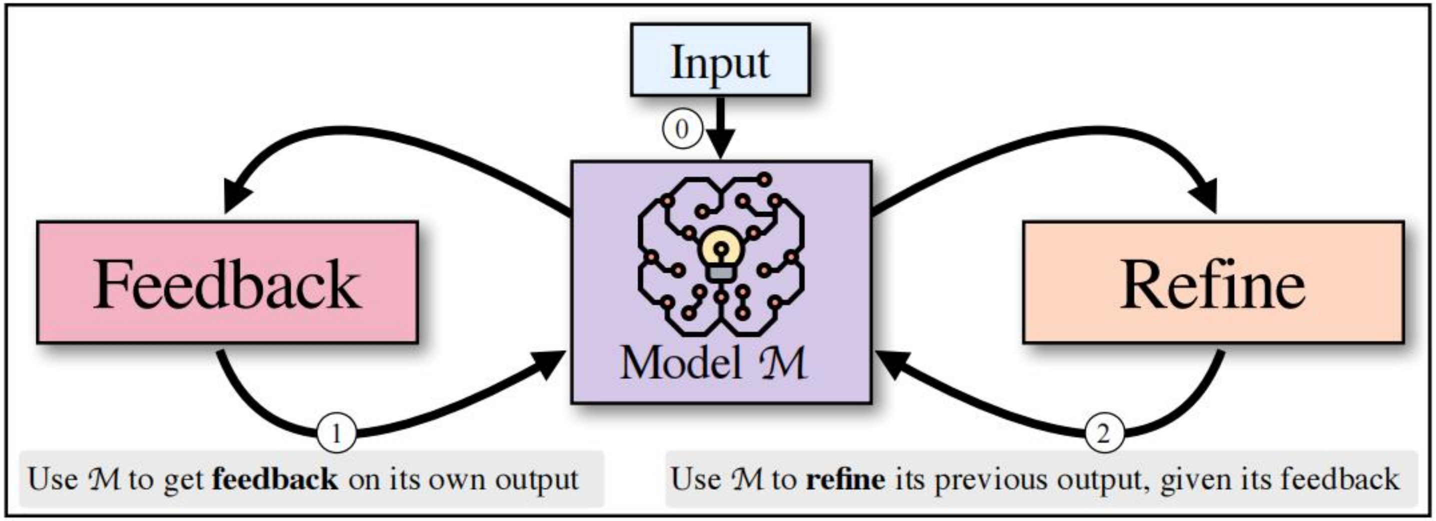
中国科学院自动化研究所
Institute of Automation
Chinese Academy of Sciences

Breaking Mental Set to Improve Reasoning through Diverse Multi-Agent Debate

Yexiang Liu, Jie Cao, Zekun Li, Ran He, Tieniu Tan

Background

- LLMs often suffer from mistakes when reasoning.
- We can use stronger model to provide feedback.
- Or utilize human supervision.
- However, effective feedback is not always obtainable.
- We need to study how to teach LLMs to self-correct.


Input
⓪
Feedback
Model $\mathcal{M}$
Refine
①
②
Use $\mathcal{M}$ to get **feedback** on its own output
Use $\mathcal{M}$ to **refine** its previous output, given its feedback

The framework of *Self-Reflection*.

Self-Refine: Iterative Refinement with Self-Feedback. NeurIPS 2023.
Language Models can Solve Computer Tasks. NeurIPS 2023.
Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS 2023.

Limitations to *Self-Reflection*

- Many studies have found limitations to Self-Reflection, performing poorly on reasoning or even worsening.
- If we provide oracles, LLMs can gradually reach the correct answer. However, in practice, we do not know the answer, neither do LLMs.

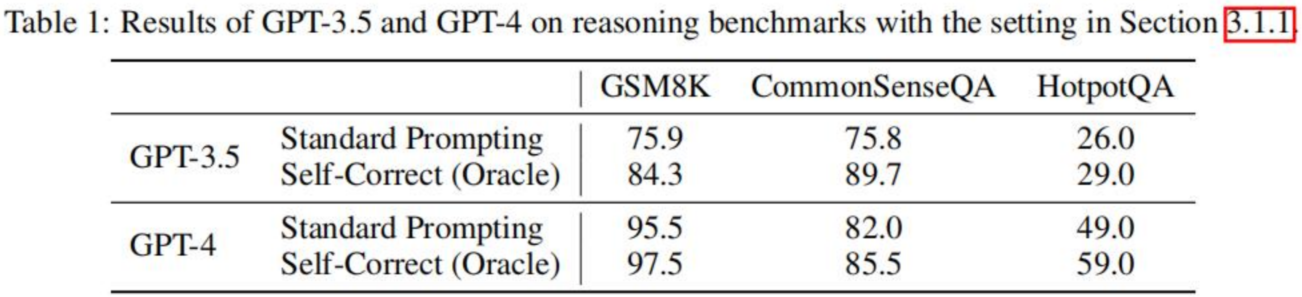


Table 1: Results of GPT-3.5 and GPT-4 on reasoning benchmarks with the setting in Section 3.1.1.

		GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	75.9	75.8	26.0
	Self-Correct (Oracle)	84.3	89.7	29.0
GPT-4	Standard Prompting	95.5	82.0	49.0
	Self-Correct (Oracle)	97.5	85.5	59.0

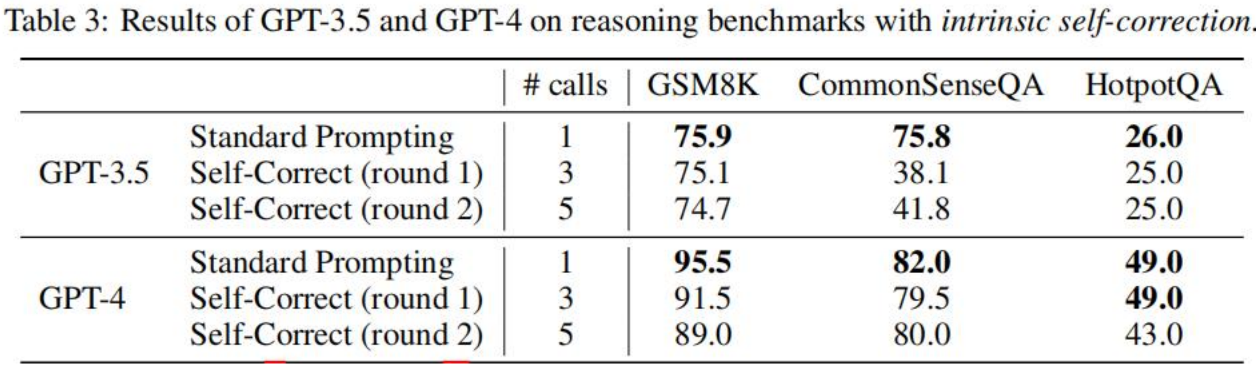


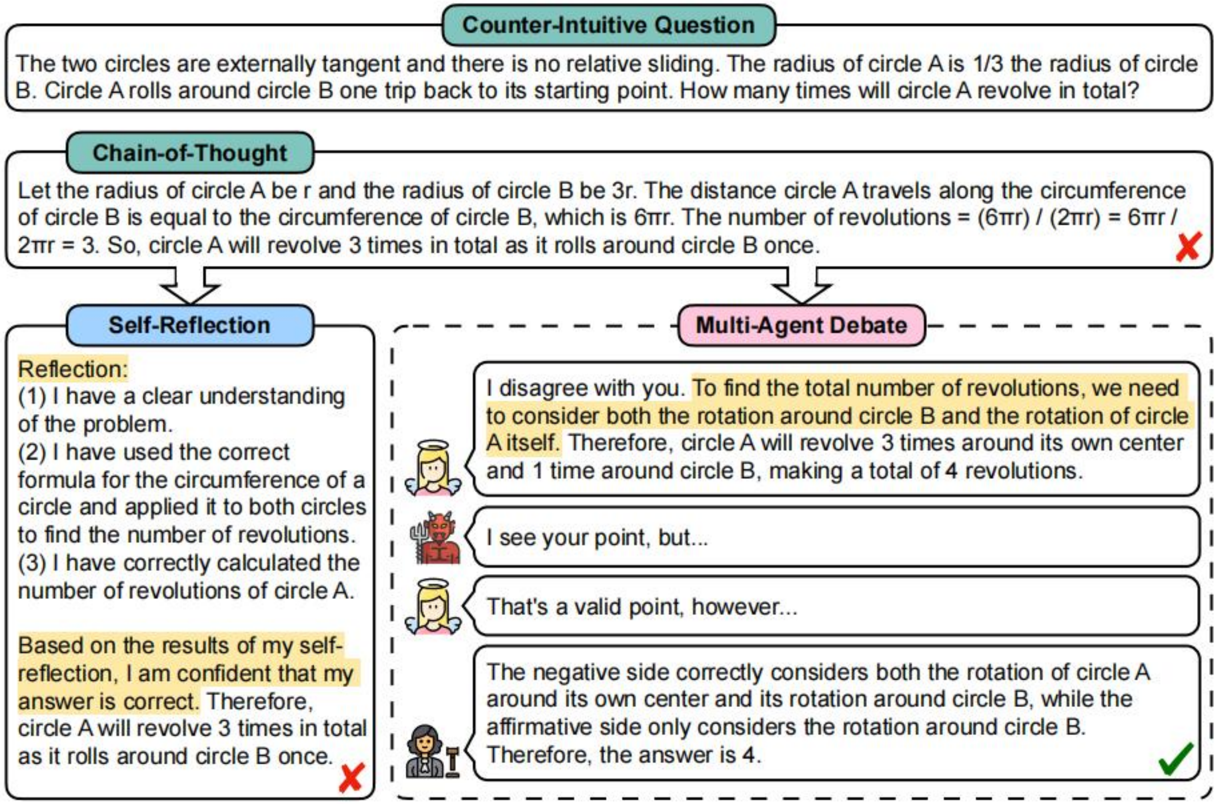
Table 3: Results of GPT-3.5 and GPT-4 on reasoning benchmarks with *intrinsic self-correction*.

		# calls	GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	1	**75.9**	**75.8**	**26.0**
	Self-Correct (round 1)	3	75.1	38.1	25.0
	Self-Correct (round 2)	5	74.7	41.8	25.0
GPT-4	Standard Prompting	1	**95.5**	**82.0**	**49.0**
	Self-Correct (round 1)	3	91.5	79.5	**49.0**
	Self-Correct (round 2)	5	89.0	80.0	43.0

Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024.
GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems. NeurIPS 2023.
Can Large Language Models Really Improve by Self-critiquing Their Own Plans? NeurIPS 2023.

Multi-Agent Debate (MAD)

- The problem of *Self-Reflection*: Degeneration of Thought.
- MAD: Multiple agents express their arguments in the state of “tit for tat” and a judge manages the debate process to obtain a final solution.



Counter-Intuitive Question

The two circles are externally tangent and there is no relative sliding. The radius of circle A is 1/3 the radius of circle B. Circle A rolls around circle B one trip back to its starting point. How many times will circle A revolve in total?

Chain-of-Thought

Let the radius of circle A be r and the radius of circle B be 3r. The distance circle A travels along the circumference of circle B is equal to the circumference of circle B, which is 6πr. The number of revolutions = (6πr) / (2πr) = 6πr / 2πr = 3. So, circle A will revolve 3 times in total as it rolls around circle B once. ✗

Self-Reflection

Reflection:
(1) I have a clear understanding of the problem.
(2) I have used the correct formula for the circumference of a circle and applied it to both circles to find the number of revolutions.
(3) I have correctly calculated the number of revolutions of circle A.

Based on the results of my self-reflection, I am confident that my answer is correct. Therefore, circle A will revolve 3 times in total as it rolls around circle B once. ✗

Multi-Agent Debate

I disagree with you. To find the total number of revolutions, we need to consider both the rotation around circle B and the rotation of circle A itself. Therefore, circle A will revolve 3 times around its own center and 1 time around circle B, making a total of 4 revolutions.

I see your point, but...

That's a valid point, however...

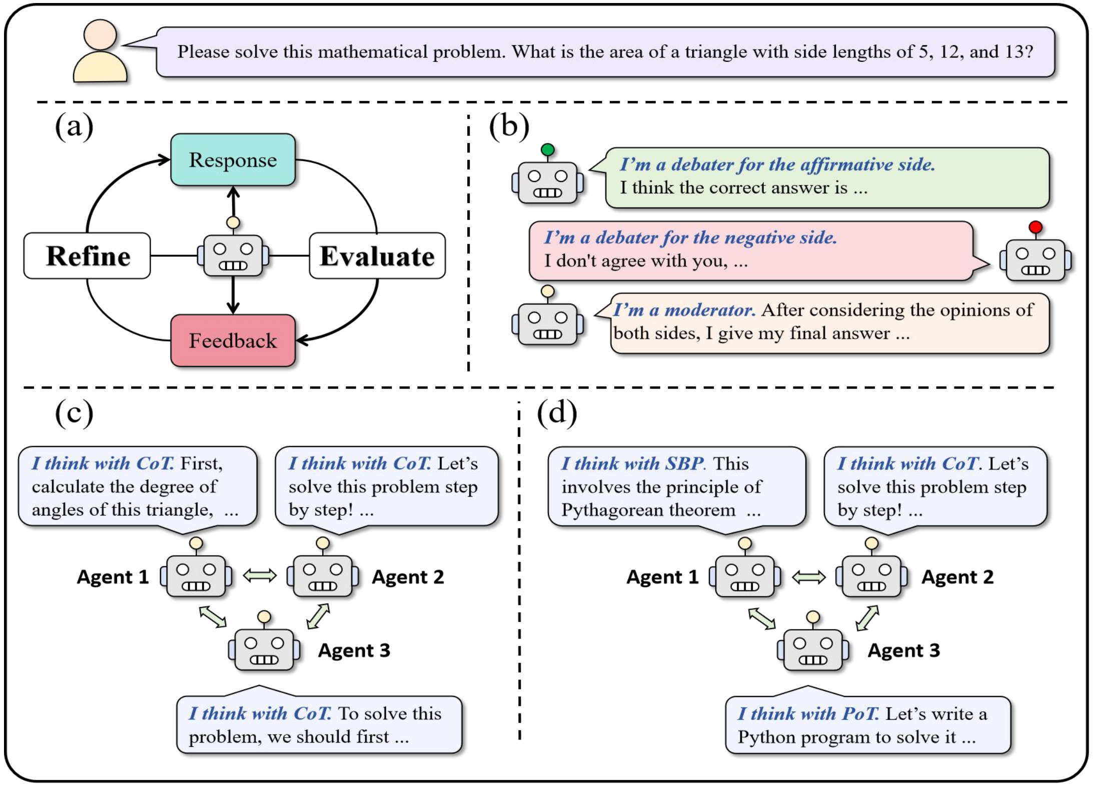
The negative side correctly considers both the rotation of circle A around its own center and its rotation around circle B, while the affirmative side only considers the rotation around circle B. Therefore, the answer is 4. ✓

Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. EMNLP 2024.
Improving Factuality and Reasoning in Language Models through Multiagent Debate. ICML 2024.

Diverse Multi-Agent Debate (DMAD)

- Both *Self-Reflection* and MAD suffer from the mental set.
- We propose DMAD. By leveraging diverse problem-solving strategies, each agent can gain insights from different perspectives, refining its responses through discussion and collectively arriving at the optimal solution.

(a) *Self-Reflection*
(b) MAD-persona
(c) MAD
(d) DMAD


Please solve this mathematical problem. What is the area of a triangle with side lengths of 5, 12, and 13?

I'm a debater for the affirmative side. I think the correct answer is ...

I'm a debater for the negative side. I don't agree with you, ...

I'm a moderator. After considering the opinions of both sides, I give my final answer ...

I think with CoT. First, calculate the degree of angles of this triangle, ...

I think with CoT. Let's solve this problem step by step! ...

I think with CoT. To solve this problem, we should first ...

I think with SBP. This involves the principle of Pythagorean theorem ...

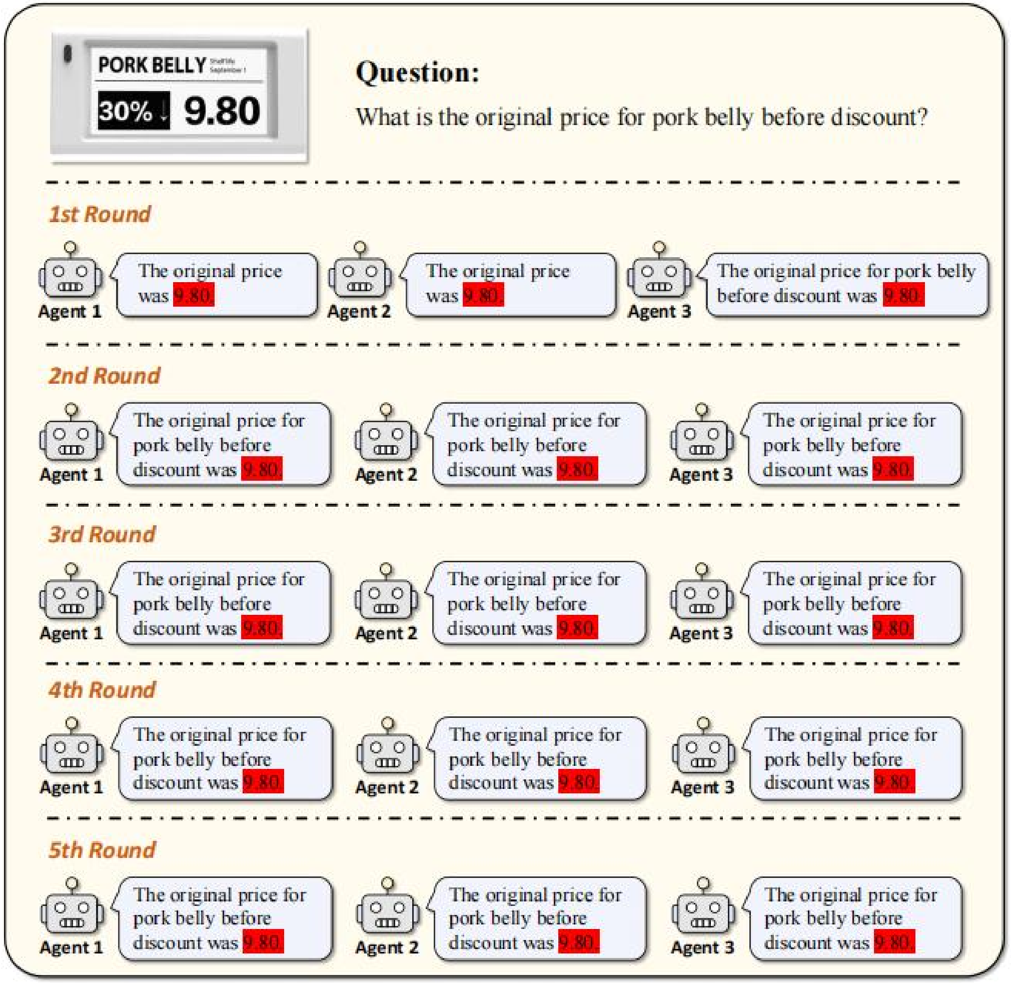
I think with CoT. Let's solve this problem step by step! ...

I think with PoT. Let's write a Python program to solve it ...

Breaking Mental Set to Improve Reasoning through Diverse Multi-Agent Debate. ICLR 2025.

Diverse Multi-Agent Debate (DMAD)

- MAD with a fixed strategy may always get the wrong answer.
- However, it succeeds just by transforming its thinking.



Question:

What is the original price for pork belly before discount?

1st Round

Agent 1: The original price was 9.80

Agent 2: The original price was 9.80

Agent 3: The original price for pork belly before discount was 9.80

2nd Round

Agent 1: The original price for pork belly before discount was 9.80

Agent 2: The original price for pork belly before discount was 9.80

Agent 3: The original price for pork belly before discount was 9.80

3rd Round

Agent 1: The original price for pork belly before discount was 9.80

Agent 2: The original price for pork belly before discount was 9.80

Agent 3: The original price for pork belly before discount was 9.80

4th Round

Agent 1: The original price for pork belly before discount was 9.80

Agent 2: The original price for pork belly before discount was 9.80

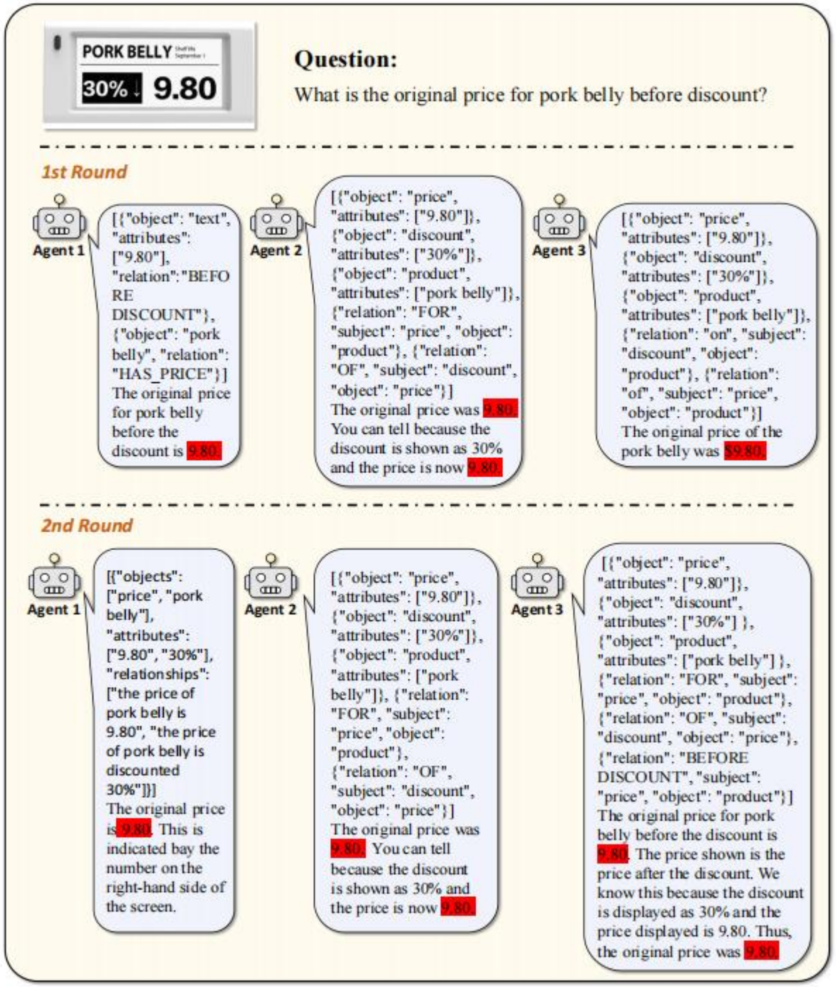
Agent 3: The original price for pork belly before discount was 9.80

5th Round

Agent 1: The original price for pork belly before discount was 9.80

Agent 2: The original price for pork belly before discount was 9.80

Agent 3: The original price for pork belly before discount was 9.80



Question:

What is the original price for pork belly before discount?

1st Round

Agent 1: [{"object": "text", "attributes": ["9.80"], "relation":"BEFORE DISCOUNT"}, {"object": "pork belly", "relation": "HAS_PRICE"}] The original price for pork belly before the discount is 9.80

Agent 2: [{"object": "price", "attributes": ["9.80"]}, {"object": "discount", "attributes": ["30%"]}, {"object": "product", "attributes": ["pork belly"]}, {"relation": "FOR", "subject": "price", "object": "product"}, {"relation": "OF", "subject": "discount", "object": "price"}] The original price was 9.80. You can tell because the discount is shown as 30% and the price is now 9.80.

Agent 3: [{"object": "price", "attributes": ["9.80"]}, {"object": "discount", "attributes": ["30%"]}, {"object": "product", "attributes": ["pork belly"]}, {"relation": "on", "subject": "discount", "object": "product"}, {"relation": "of", "subject": "price", "object": "product"}] The original price of the pork belly was $9.80.

2nd Round

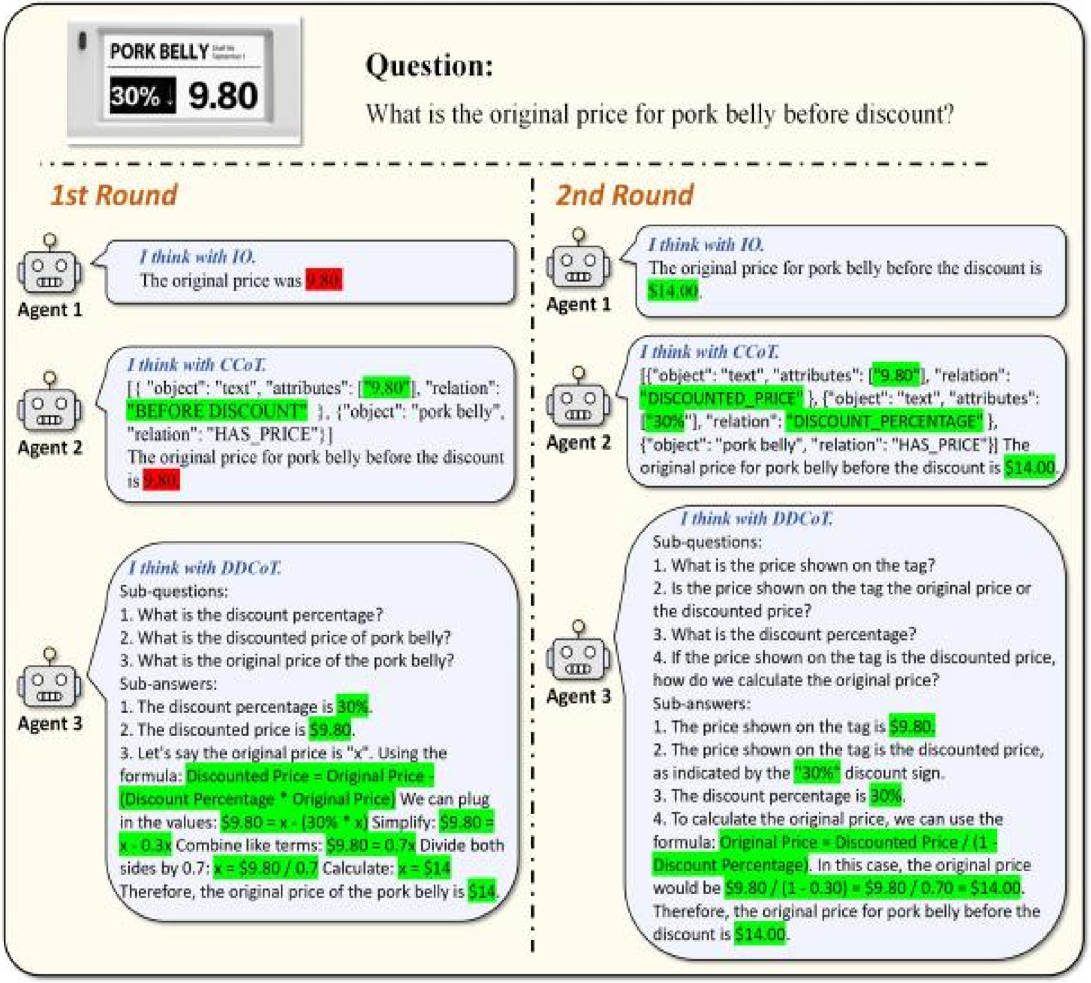
Agent 1: [{"objects": ["price", "pork belly"], "attributes": ["9.80", "30%"], "relationships": ["the price of pork belly is 9.80", "the price of pork belly is discounted 30%"]}] The original price is 9.80. This is indicated bay the number on the right-hand side of the screen.

Agent 2: [{"object": "price", "attributes": ["9.80"]}, {"object": "discount", "attributes": ["30%"]}, {"object": "product", "attributes": ["pork belly"]}, {"relation": "FOR", "subject": "price", "object": "product"}, {"relation": "OF", "subject": "discount", "object": "price"}] The original price was 9.80. You can tell because the discount is shown as 30% and the price is now 9.80.

Agent 3: [{"object": "price", "attributes": ["9.80"]}, {"object": "discount", "attributes": ["30%"] }, {"object": "product", "attributes": ["pork belly"] }, {"relation": "FOR", "subject": "price", "object": "product"}, {"relation": "OF", "subject": "discount", "object": "price"}, {"relation": "BEFORE DISCOUNT", "subject": "price", "object": "product"}] The original price for pork belly before the discount is 9.80. The price shown is the price after the discount. We know this because the discount is displayed as 30% and the price displayed is 9.80. Thus, the original price was 9.80.

Diverse Multi-Agent Debate (DMAD)

- MAD with a fixed strategy may always get the wrong answer.
- However, it succeeds just by transforming its thinking.



Question:

What is the original price for pork belly before discount?

1st Round

Agent 1

I think with IO.
The original price was 9.80

Agent 2

I think with CCoT.
[{ "object": "text", "attributes": ["9.80"], "relation": "BEFORE DISCOUNT" }, {"object": "pork belly", "relation": "HAS_PRICE"}]
The original price for pork belly before the discount is 9.80

Agent 3

I think with DDCoT.
Sub-questions:
1. What is the discount percentage?
2. What is the discounted price of pork belly?
3. What is the original price of the pork belly?

Sub-answers:
1. The discount percentage is 30%.
2. The discounted price is $9.80.
3. Let's say the original price is "x". Using the formula: Discounted Price = Original Price - (Discount Percentage * Original Price) We can plug in the values: $9.80 = x - (30% * x) Simplify: $9.80 = x - 0.3x Combine like terms: $9.80 = 0.7x Divide both sides by 0.7: x = $9.80 / 0.7 Calculate: x = $14

Therefore, the original price of the pork belly is $14.

2nd Round

Agent 1

I think with IO.
The original price for pork belly before the discount is $14.00.

Agent 2

I think with CCoT.
[{"object": "text", "attributes": ["9.80"], "relation": "DISCOUNTED_PRICE" }, {"object": "text", "attributes": ["30%"], "relation": "DISCOUNT_PERCENTAGE" }, {"object": "pork belly", "relation": "HAS_PRICE"}] The original price for pork belly before the discount is $14.00.

Agent 3

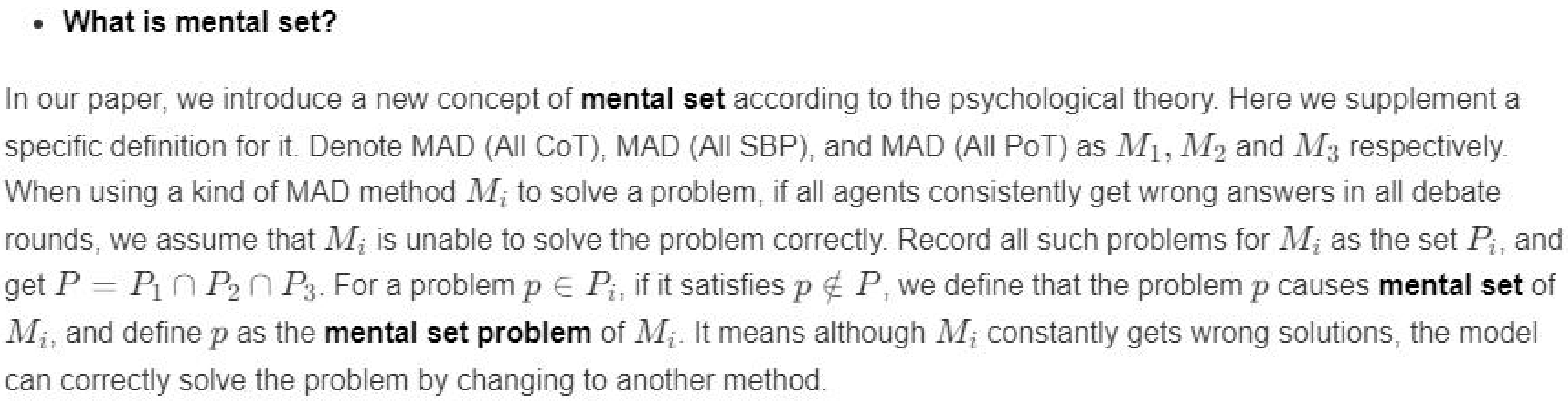
I think with DDCoT.
Sub-questions:
1. What is the price shown on the tag?
2. Is the price shown on the tag the original price or the discounted price?
3. What is the discount percentage?
4. If the price shown on the tag is the discounted price, how do we calculate the original price?

Sub-answers:
1. The price shown on the tag is $9.80.
2. The price shown on the tag is the discounted price, as indicated by the "30%" discount sign.
3. The discount percentage is 30%.
4. To calculate the original price, we can use the formula: Original Price = Discounted Price / (1 - Discount Percentage). In this case, the original price would be $9.80 / (1 - 0.30) = $9.80 / 0.70 = $14.00.

Therefore, the original price for pork belly before the discount is $14.00.

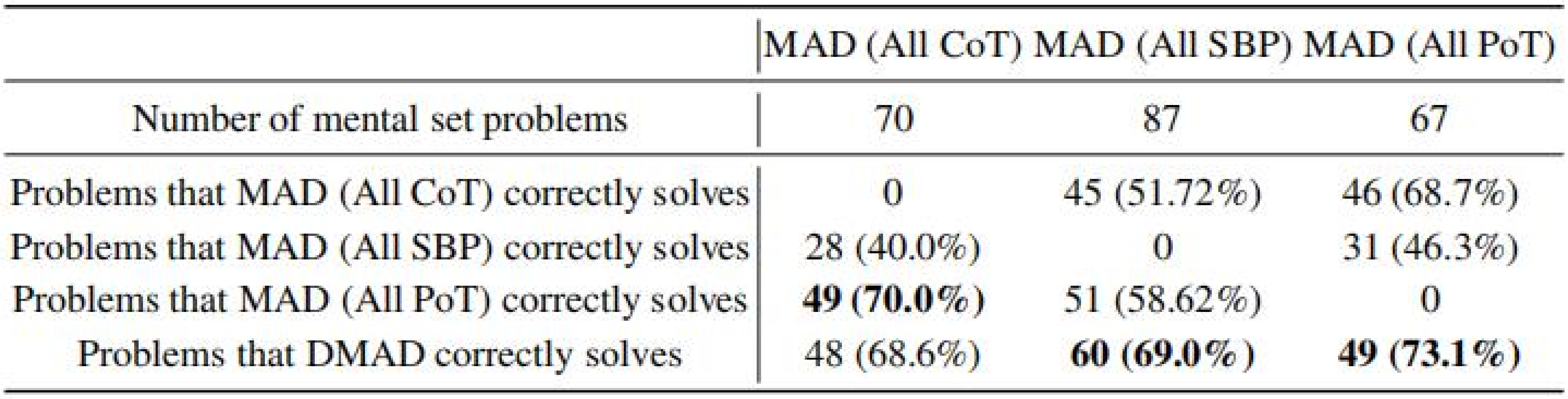
Diverse Multi-Agent Debate (DMAD)

- DMAD can more effectively solve other methods’ mental set problems.



- **What is mental set?**

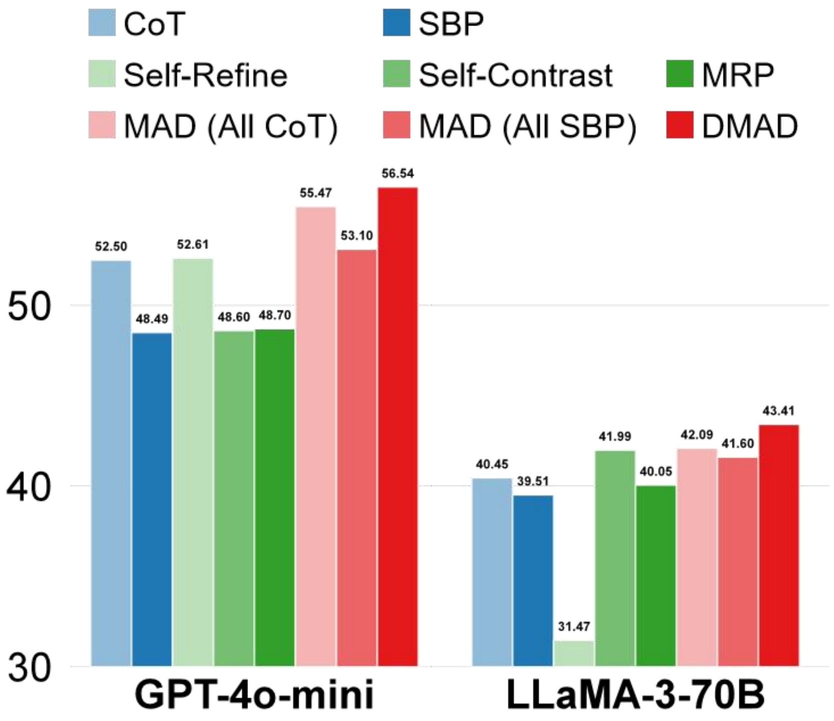
In our paper, we introduce a new concept of **mental set** according to the psychological theory. Here we supplement a specific definition for it. Denote MAD (All CoT), MAD (All SBP), and MAD (All PoT) as M_1, M_2 and M_3 respectively. When using a kind of MAD method M_i to solve a problem, if all agents consistently get wrong answers in all debate rounds, we assume that M_i is unable to solve the problem correctly. Record all such problems for M_i as the set P_i, and get $P = P_1 \cap P_2 \cap P_3$. For a problem $p \in P_i$, if it satisfies $p \notin P$, we define that the problem p causes **mental set** of M_i, and define p as the **mental set problem** of M_i. It means although M_i constantly gets wrong solutions, the model can correctly solve the problem by changing to another method.



	MAD (All CoT)	MAD (All SBP)	MAD (All PoT)
Number of mental set problems	70	87	67
Problems that MAD (All CoT) correctly solves	0	45 (51.72%)	46 (68.7%)
Problems that MAD (All SBP) correctly solves	28 (40.0%)	0	31 (46.3%)
Problems that MAD (All PoT) correctly solves	**49 (70.0%)**	51 (58.62%)	0
Problems that DMAD correctly solves	48 (68.6%)	**60 (69.0%)**	**49 (73.1%)**

Diverse Multi-Agent Debate (DMAD)

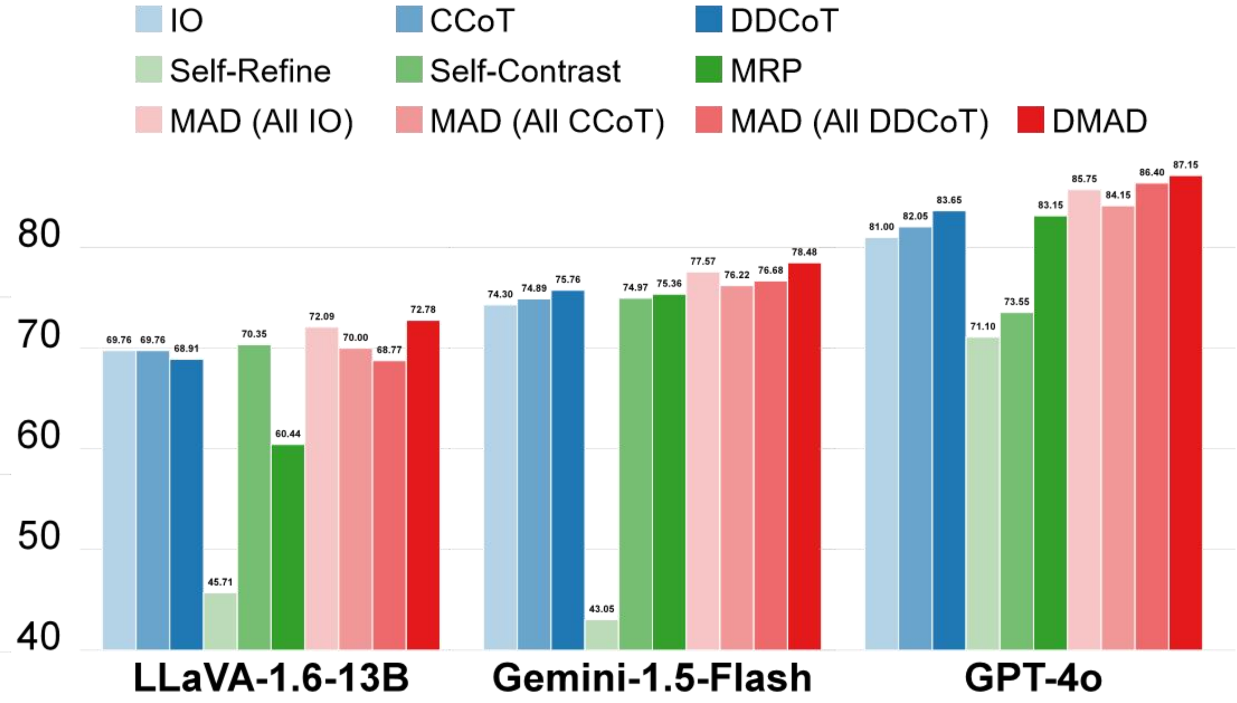
- We evaluate DMAD against various prompting techniques, including *self-reflection* and traditional MAD, across multiple benchmarks including math, chemistry, physics, biology and so on, using both LLMs and Multimodal LLMs. Our experiments show that DMAD consistently outperforms other methods.



(a) LLMs

Diverse Multi-Agent Debate (DMAD)

- We evaluate DMAD against various prompting techniques, including *self-reflection* and traditional MAD, across multiple benchmarks including math, chemistry, physics, biology and so on, using both LLMs and Multimodal LLMs. Our experiments show that DMAD consistently outperforms other methods.



(b) MLLMs

Diverse Multi-Agent Debate (DMAD)

- What’s more, DMAD can deliver better results than MAD in fewer rounds, and perform better when increasing the number of reasoning methods (for n = 1,2,3.).

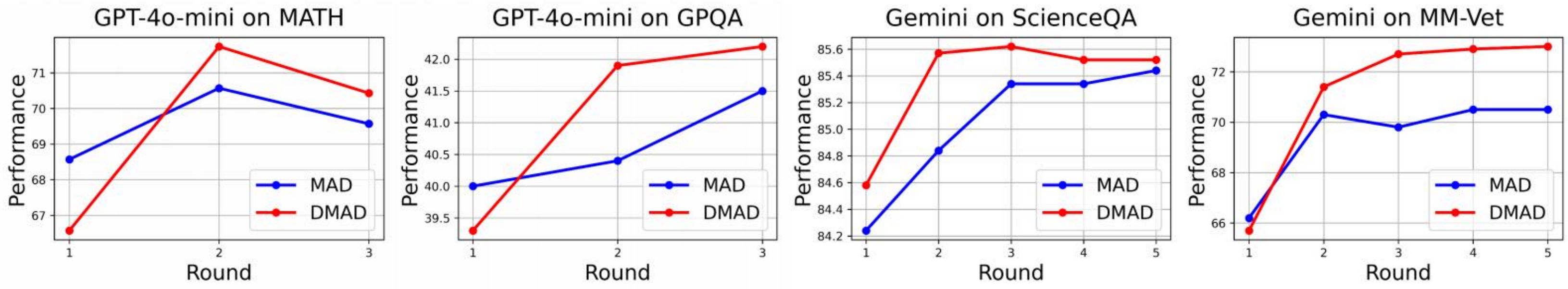


Figure 4: Performance with increased rounds. More results are shown in Figure 6.

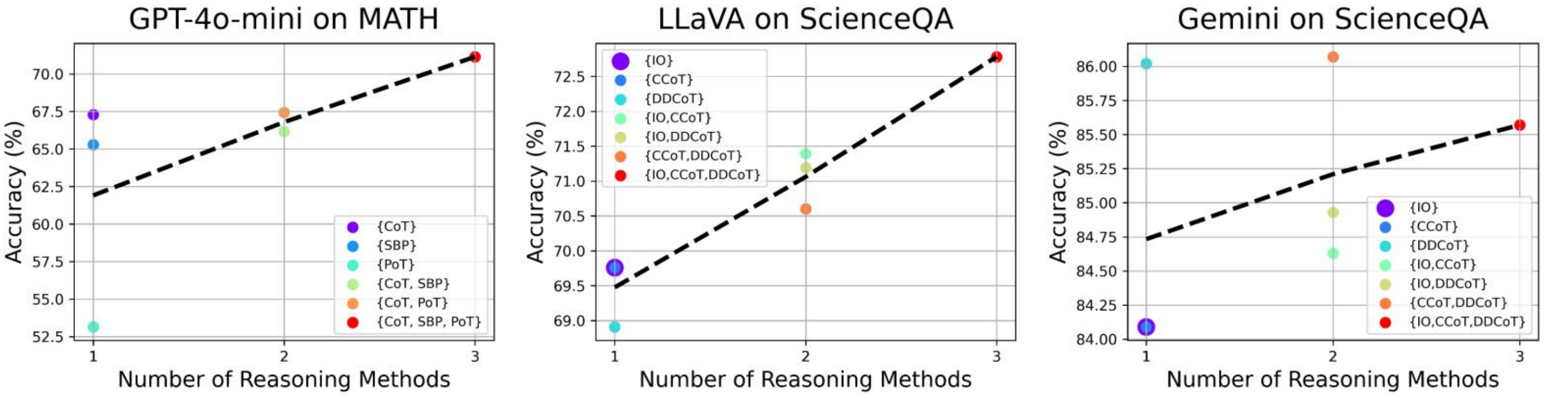


Figure 5: Performance vs the number of reasoning methods on DMAD.

Limitations and Future work

■ Although DMAD can perform better, there are still some limitations:

- The inherent problem of MAD: LLM agents are easily influenced by other agents that come to incorrect conclusions, mistakenly changing their original correct solutions to wrong ones.
- There is still room for improvement in accuracy.
- How to design an intelligent system where agents can dynamically select the most suitable reasoning strategy.
- High overhead.

Thanks