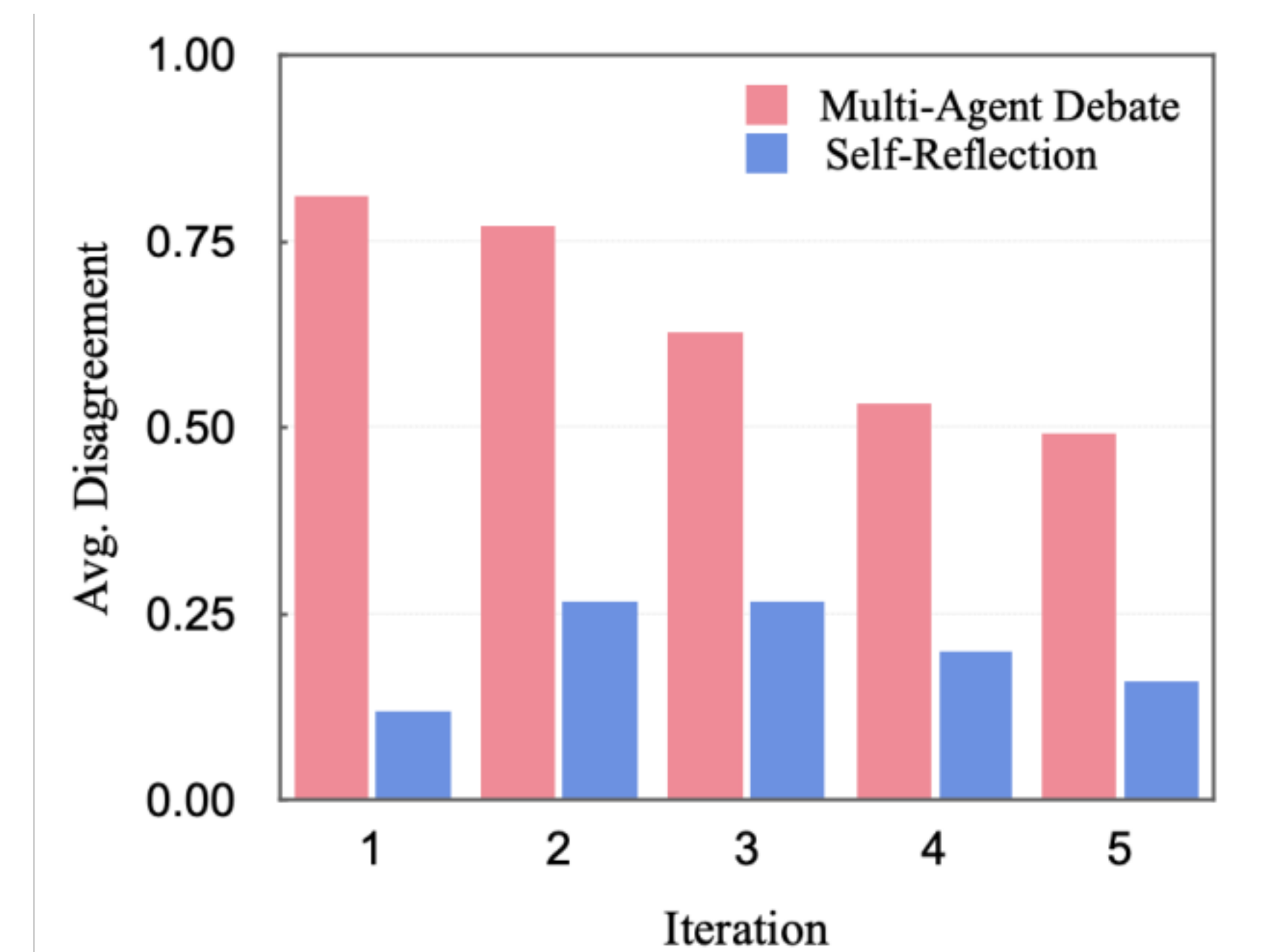
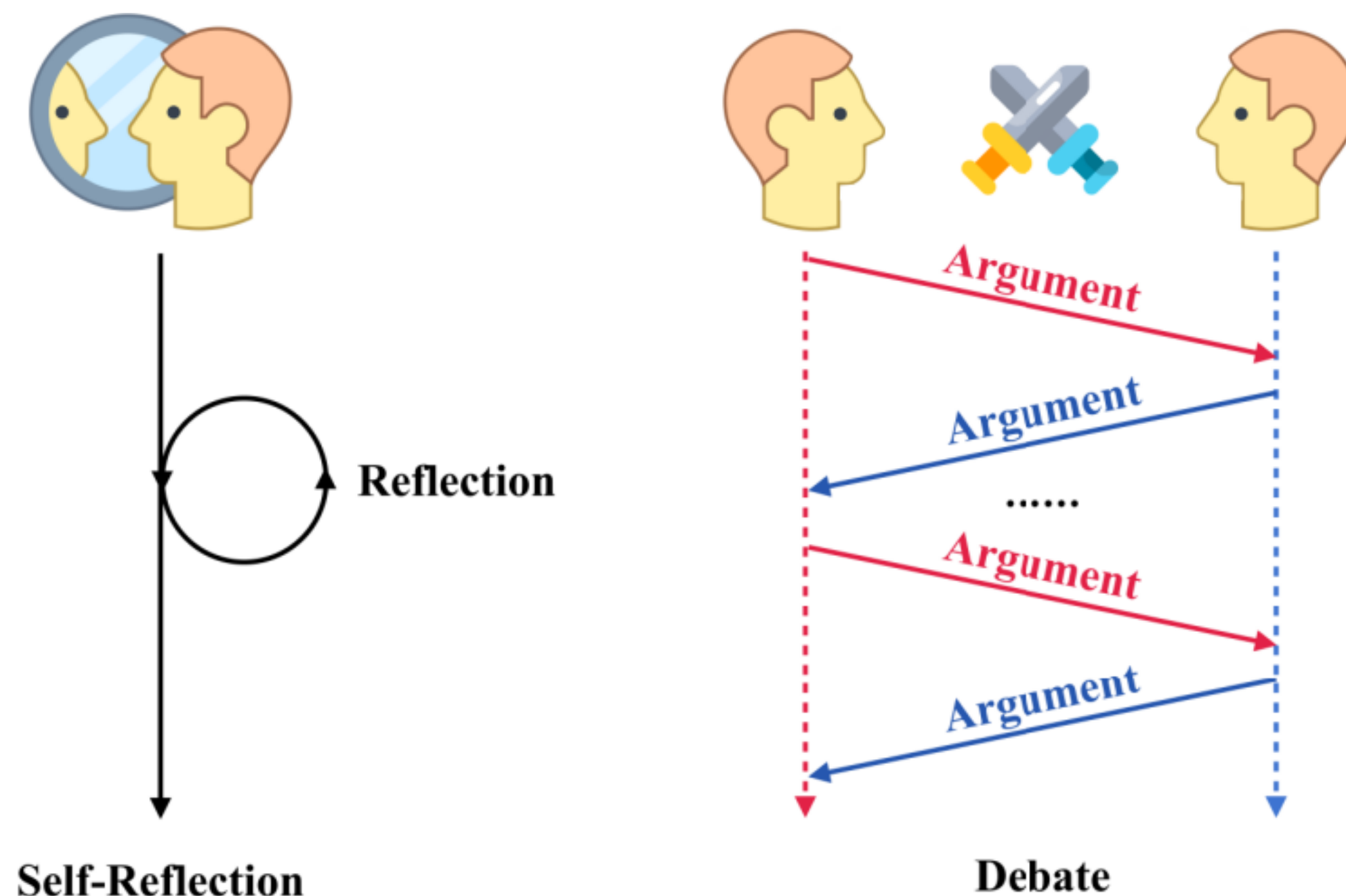


Limitations of Current LLMs

Limitation: Degeneration-of-Thought (DoT)

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



Encouraging Divergent Thinking in LLMs

MAD: Multi-Agent Debate

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 $6\pi r$. The number of revolutions = $(6\pi r) / (2\pi r) = 6\pi r / 2\pi r = 3$. 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. ✅

Algorithm

Algorithm 1 MAD: Multi-Agent Debate


Require: Debate topic t , maximum number of rounds M and number of debaters N

Ensure: Final answer a


```
1: procedure MAD( $t, M, N$ )
2:    $J$   $\triangleright$  Initialize the judge
3:    $D \leftarrow [D_1, \dots, D_N]$   $\triangleright$  Initialize debaters
4:    $H \leftarrow [t]$   $\triangleright$  Initialize debate history
5:    $m \leftarrow 0$   $\triangleright$  Current round
6:   while  $m \leq M$  do
7:      $m \leftarrow m + 1$ 
8:     for each  $D_i$  in  $D$  do
9:        $h \leftarrow D_i(H)$   $\triangleright$  Generate argument
10:       $H \leftarrow H + [h]$   $\triangleright$  Append  $h$  to  $H$ 
11:     if  $J_d(H)$  then
12:       break  $\triangleright$  Debate is over
13:    $a \leftarrow J_e(H)$   $\triangleright$  Extract the final answer
14:   return  $a$ 
```

Debater Public Prompt ( & )

You are a debater. Hello and welcome to the debate competition. It's not necessary to fully agree with each other's perspectives, as our objective is to find the correct answer. The debate topic is stated as follows: <debate topic>.

Affirmative Private Prompt ()

You are affirmative side. Please express your viewpoints.

Negative Private Prompt ()

You are negative side. You disagree with the affirmative side's points. Provide your reasons and answer.

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.

Experimental Setup

Task 1: Commonsense Machine Translation

Ambiguity Type	Source Sentence	Correct Reference	Incorrect Translation
Lexical	<u>吃掉</u> 敌人一个师。	<u>Destroy</u> a division of the enemy.	<u>Eat up</u> an enemy division.
	他喜欢吃 <u>吃</u> 苹果。	He likes to <u>eat</u> apples.	He likes to <u>destory</u> apples.
Contextless	<u>正在手术的</u> 是健康的医生。	A healthy doctor <u>is doing surgery</u> .	What <u>is undergoing surgery</u> is a doctor who is healthy.
	<u>正在手术的</u> 是生命垂危的病人。	What <u>is undergoing surgery</u> is a patient whose life is dying.	A patient whose life is dying <u>is doing surgery</u> .
Contextual	当地震袭击中国时， <u>援助的是中国</u> 。	When the earthquake hit China, <u>China was aided</u> .	When the earthquake hit China, <u>China has assisted</u> .
	当地震袭击日本时， <u>援助的是中国</u> 。	When the earthquake hit Japan, <u>China has assisted</u> .	When the earthquake hit Japan, <u>China was aided</u> .

Experimental Results

Task 1: Commonsense Machine Translation

Method	Lexical			Contextless			Contextual		
	COMET	BLEURT	SCORE	COMET	BLEURT	SCORE	COMET	BLEURT	SCORE
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 ↑ 0.02	84.5	73.2	3.46 ↑ 0.03	85.3	73.9	3.58 ↑ 0.01
+ MAPS	81.9	70.1	3.43 ↑ 0.29	84.2	73.5	3.45 ↑ 0.02	85.2	74.0	3.56 ↓ 0.01
+ Self-Reflect	81.0	69.1	3.43 ↑ 0.29	83.6	72.2	3.46 ↑ 0.03	84.9	73.5	3.63 ↑ 0.06
+ MAD	82.0	70.9	3.78 ↑ 0.64	84.8	73.7	3.67 ↑ 0.24	85.3	74.0	3.67 ↑ 0.10
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

Experimental Setup

Task 2: Counter-Intuitive Arithmetic Reasoning

Components	Content
Question	When Alice walks up the hill, her speed is 1 m/s and when she goes down the hill, her speed is 3 m/s. Then when Alice walks up and down the hill, what is her average speed?
Correct Answer	1.5 m/s
Explanation	If Alice covers a distance of d going up and down the hill, then her total distance is $2d$. Her time going up the hill is $d/1 = d$, and her time going down the hill is $d/3$. So, her total time is $d + d/3 = 4d/3$. Therefore, her average speed is $2d / (4d/3) = 3/2$ m/s.
Incorrect Answer	2 m/s
Explanation	Alice's average speed can be calculated by adding her speed going up the hill and her speed going down the hill, and then dividing by 2. So, $(1 \text{ m/s} + 3 \text{ m/s}) / 2 = 2 \text{ m/s}$. Therefore, Alice's average speed is 2 m/s.

Experimental Results

Task 2: Counter-Intuitive Arithmetic Reasoning

Question A	
The two circles are externally tangent and there is no relative sliding. The radius of circle A is $\frac{1}{3}$ 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?	
Correct Answer	4
GPT-4	4
GPT-3.5-Turbo	3
+ Self-Reflect	3
+ MAD	4

Question B	
When Alice walks up the hill, her speed is 1 m/s and when she goes down the hill, her speed is 3 m/s. Then when Alice walks up and down the hill, what is her average speed?	
Correct Answer	1.5 m/s
GPT-4	1.5 m/s
GPT-3.5-Turbo	2 m/s
+ Self-Reflect	2 m/s
+ MAD	1.5 m/s

Method	ACC (%)
Human	60.0
GPT-4	51.0
GPT-3.5-Turbo	26.0
+ CoT	28.0
+ Self-Consistency	29.5
+ Self-Reflect	27.5
+ MAD	37.0

- Self-Reflect does not improve much over the baseline GPT-3.5-Turbo
- MAD outperforms all the other compared methods

Mitigation of DoT

① Bias and Distorted Perception

- Bias: Ambiguity error rate (human evaluation)

② Rigidity and Resistance to Change

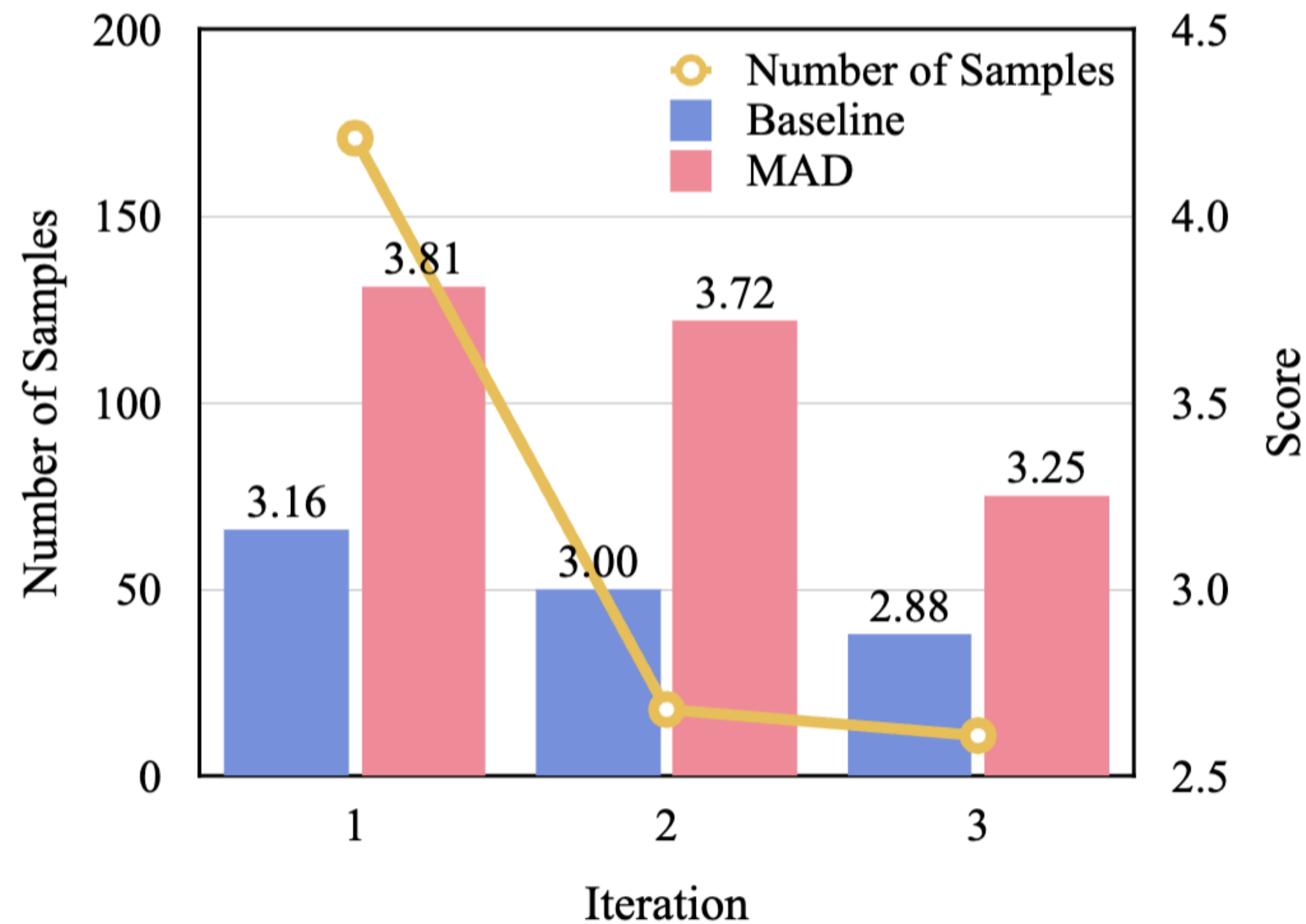
- Diversity: Self-BLEU score

③ Limited External Feedback.

- We introduce the views of other agents in the form of debates

Method	Lexical		Contextless		Contextual	
	Bias↓	Diversity↑	Bias↓	Diversity↑	Bias↓	Diversity↑
Self-Reflect	31.0	24.5	22.2	17.3	36.6	18.8
MAD	20.5	48.7	18.2	61.0	35.7	35.8

Iteration Distribution of Debate



- Efficiency
 - ✓ In the majority of cases, the optimal answer can be achieved through a single round of debate
- Effectiveness
 - ✓ When translating more complex sentences, the judge requires additional iterations to gather adequate information

Weak LLM as Judge







- The quality of the debaters' responses significantly impact the performance ceiling of MAD.
- The selection of the judge agent plays a secondary role.

Method	Lexical			Contextless			Contextual		
	COMET	BLEURT	SCORE	COMET	BLEURT	SCORE	COMET	BLEURT	SCORE
GPT-3.5-Turbo as Debaters									
Turbo as Judge	82.0	70.9	3.78	84.8	73.7	3.67	85.3	74.0	3.67
Vicuna as Judge	81.2	69.1	3.46	83.1	70.3	3.41	84.4	73.1	3.55
Vicuna-13b as Debaters									
Turbo as Judge	76.7	64.8	2.96	79.7	67.6	3.02	82.1	70.6	3.43
Vicuna as Judge	77.2	65.1	2.96	80.1	67.3	3.11	82.6	70.9	3.45

Case Study

Task 1: Commonsense Machine Translation







Source	吃掉敌人一个师。
Correct Reference	Destroy a division of the enemy.
Incorrect Reference	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.

-  I believe the correct translation of the Chinese text “吃掉敌人一个师” is “Eat up an enemy division.” The word “吃掉” literally means “eat up” or “consume.”
-  I disagree with your translation. I believe the correct translation of the Chinese text “吃掉敌人一个师” is “Eliminate an enemy division.” While “吃掉” can be translated as “eat up” or “consume”, it can also mean “eliminate” or “annihilate” in a military context. In this case, it is more appropriate to use “eliminate” as it conveys the idea of completely destroying the enemy division.
-  Both sides presented valid arguments and translations that accurately convey the meaning of the original Chinese text. The choice between “eat up” and “eliminate” may depend on the context and tone of the text. Therefore, there is no clear preference for either side at this time.
-  While “eliminate” does convey the idea of completely destroying the enemy division, it doesn’t capture the connotation of consuming or devouring the enemy that is implied by “吃掉”.
-  In a military context, “吃掉” is often used to describe the complete destruction of the enemy, rather than the consumption of their resources or territory.
-  Both sides presented valid arguments, but the negative side’s argument that “eliminate” better captures the intensity and seriousness of the military language used in the original Chinese text was more convincing. Additionally, “eliminate” accurately conveys the idea of completely destroying the enemy division. Therefore, the correct English translation of the Chinese text “吃掉敌人一个师” is “Eliminate an enemy division.”

Case Study

Task 2: Counter-Intuitive Arithmetic Reasoning

Question B	
When Alice walks up the hill, her speed is 1 m/s and when she goes down the hill, her speed is 3 m/s. Then when Alice walks up and down the hill, what is her average speed?	
Correct Answer	1.5 m/s
GPT-4	1.5 m/s
GPT-3.5-Turbo	2 m/s
+ Self-Reflect	2 m/s
+ MAD	1.5 m/s

-  Alice's average speed can be calculated by adding her speed going up the hill and her speed going down the hill, and then dividing by 2. So, $(1 \text{ m/s} + 3 \text{ m/s}) / 2 = 2 \text{ m/s}$. Therefore, Alice's average speed is 2 m/s.
-  I disagree with your answer. The problem doesn't give us any information about the distance Alice covers while going up and down the hill. So, we cannot simply add her speeds and divide by 2 to get her average speed. We need to know the distance she covers to calculate her average speed.
-  The average speed is calculated by dividing the total distance by the total time taken. As Alice walks up and down the hill, the distance covered in both directions is the same. Therefore, the total distance covered is 2 times the distance of either direction.
-  That's a good point. But we can assume that Alice covers the same distance going up and down the hill. In that case, we can use the formula for average speed, which is total distance divided by total time. If Alice covers a distance of d going up and down the hill, then her total distance is $2d$. Her time going up the hill is $d/1 = d$, and her time going down the hill is $d/3$. So, her total time is $d + d/3 = 4d/3$. Therefore, her average speed is $2d/(4d/3) = 3/2 \text{ m/s}$.
-  Thank you for providing your answer. This average speed of 1.5 m/s takes into account both the uphill and downhill speeds, providing a balanced representation of Alice's overall performance while walking up and down the hill.
-  Yes. Let's assume the distance is d . The time taken to walk up the hill is $d/1$ and the time taken to walk down the hill is $d/3$. The total time taken is $d/1 + d/3 = 4d/3$. Therefore, the average speed is $(2d)/(4d/3) = 1.5 \text{ m/s}$. Hence, the negative side is correct.

Summary

- ☑ LLM can improve translation quality by mimicking human translation strategies.
- ☑ MT model can learn from human feedback (modeled by QE) after addressing overoptimization.
- ☑ Multi-Agent Debate can address Degeneration-of-Thought (DoT) problem by exploring divergent thinking.