

Diverse Thinking: Breaking Oppositions in Debates to Foster Synergized Reasoning via Multi-Agent Collaboration

Zhaokang Huang^{1*}, Yiming Xu^{1*}, Tao Wang¹, Yu Yang², Tianyong Hao^{1†}

¹ School of Computer Science, South China Normal University, China

² Centre for Learning, Teaching and Technology,
The Education University of Hong Kong, China

*Equal contribution, †Correspondence: haoty@scnu.edu.cn

Abstract

Large Language Models (LLMs) frequently face limitations, such as judgment biases and cognitive deadlocks, in complex reasoning due to “Degradation of Thought” in self-reflection. While Multi-Agent Debate (MAD) approaches attempt to address this issue, their inherently adversarial nature suppresses diverse perspectives by overemphasizing antagonism, which results in the loss of partially valid reasoning. In this paper, we propose Multi-Agent Diverse Thinking (MADT), a novel cognitive framework that redefines multi-agent collaboration by shifting from adversarial confrontation toward constructive synergy. To break the oppositional deadlock found in traditional debates, MADT decomposes the thinking process into fine-grained modules: an Affirmative Thinker to preserve rational components, a Critical Thinker to rectify errors, and a Growth-Minded Thinker to provide optimization suggestions. These agents jointly foster a synergistic environment where specialized roles protect valid logic while iteratively refining flaws. Meanwhile, a Leader agent coordinates the collaborations and synthesizes the feedback from multiple thinkers to produce the final result. Extensive experiments on Common Machine Translation and Counter-Intuitive Arithmetic Reasoning tasks show that MADT consistently outperforms Self-Reflection and MAD baselines, validating the superior ability to enhance complex logical and reasoning performance in LLMs.

1 Introduction

While large language models (LLMs) have achieved remarkable results in general linguistic tasks, they exhibit poor performance in complex mathematical and logical reasoning (Bang et al., 2023; Tran et al., 2025). To address these challenges, researchers have gradually drawn inspiration from human thinking patterns and the efficient operational mechanisms of the human society to

pursue improvements to existing LLMs (Gou et al., 2023; Zhu et al., 2023).

In current research, investigators have attained phased successes. Among these advancements: Chain-of-Thought (CoT) enables models to solve problems more accurately through step-by-step reasoning, reducing the likelihood of leaping directly to erroneous conclusions (Wei et al., 2022). Self-Consistency mitigates hallucinations generated by LLMs by producing multiple candidate answers and selecting the most consistent one (Wang et al., 2022). CoVe significantly diminishes model hallucinations and enhances the factual accuracy of generated content via verification steps (Dhuliawala et al., 2024). Multi-Agent Debate (MAD) that improves LLMs’ reasoning capabilities and reduces hallucinations through debates among multiple agents (Liang et al., 2024). D-MAD, an enhanced MAD framework, breaks through fixed mental sets by encouraging agents to adopt diverse reasoning approaches (Liu et al., 2025). The above-mentioned frameworks have achieved certain efficacy; however, the rationality of the frameworks remains a direction that merits further inquiry.

Essentially, the current methods can be categorized into two major types according to their corresponding thinking patterns:

- **Self-Reflection** (Madaan et al., 2023): Such methods allow LLMs to conduct introspection and examination of their own reasoning processes, iteratively generating new answers based on previous responses and feedback. However, they also suffer from Degradation of Thought (DoT) problem — when LLMs become certain of their own answers, Self-Reflection falls into a local impasse and fails to generate new lines of thinking.
- **MAD**: This type of methods addresses the DoT problem inherent in Self-Reflection through Debate mechanisms, achieving

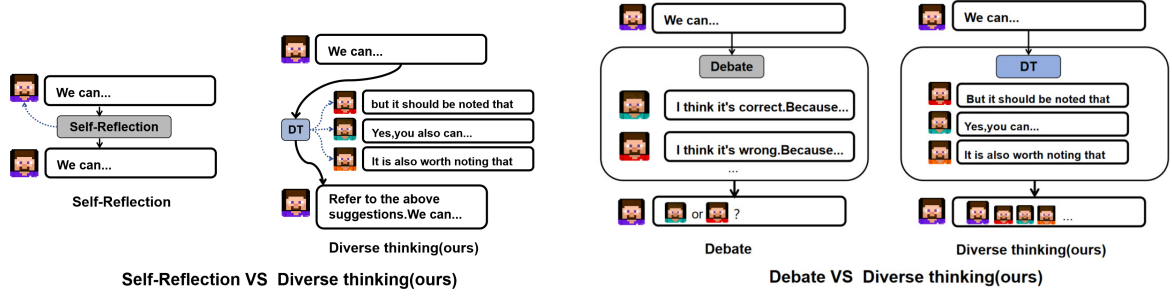


Figure 1: Comparison of Self-Reflection, Debate, and our proposed Diverse Thinking (DT)

promising results. However, the adversarial nature of debates may lead to complete overshadowing of partially reasonable answers.

In this work, we conduct an in-depth analysis of two representative cognitive paradigms—self-reflection and debate—where self-reflection suffers from DoT problem and thus requires the integration of external perspectives, while Multi-Agent Debate (MAD) overemphasizes adversariality to generate diverse viewpoints, thereby leading to the loss of partial valid reasoning steps. Oppositions are an innate, pre-existing characteristic of debate scenarios, which imposes novel requirements on multi-agent collaboration. We define multi-agent collaboration by shifting from adversarial confrontation to constructive synergy. Accordingly, we propose Diverse Thinking, a novel cognitive paradigm integrating dual advantages of external perspective incorporation and valid reasoning preservation. This paradigm abandons the preset binary premise of "right versus wrong" and fosters synergized reasoning via differentiated division of labor and collaboration.

On this basis, we design the Multi-Agent Diverse Thinking (MADT) framework, which deploys three specialized thinkers: a Growth-Minded Thinker, an Affirmative Thinker, and a Critical Thinker. The Growth-Minded Thinker is designed to address the DoT challenge in self-reflection, generating targeted improvement suggestions for initial responses and leveraging external viewpoints to mitigate the intrinsic biases of self-reflection. To resolve the adversarial dilemma in Multi-Agent Debate (MAD), we move beyond the paradigm of opposing viewpoints in traditional debates and instead focus on iterative deliberations centered on the initial answer; for this purpose, we develop an Affirmative Thinker and a Critical Thinker, which respectively preserve the rational components of

the initial answer and rectify existing flaws. Meanwhile, a Leader agent coordinates the collaborations and synthesizes the feedback from multiple thinkers to produce the final result. Figure 1 provides a comparative analysis of the framework proposed in this paper against the other two frameworks.

Our contributions are summarized as follows:

- We identify a limitation of existing debate-style collaborative reasoning methods: partially correct reasoning may be discarded during adversarial critique. To address this issue, we propose Diverse Thinking (DT), a novel cognitive paradigm that shifts multi-agent collaboration from adversarial confrontation to constructive synergy, enabling external perspective integration while preserving partially valid reasoning.
- Based on this paradigm, we develop the Multi-Agent Diverse Thinking (MADT) framework, a lightweight role-specialized collaborative refinement framework that decomposes refinement into affirmative, critical, and improvement-oriented feedback. We further introduce an affirmative refinement mechanism that explicitly preserves valid reasoning components during collaborative reasoning.
- Extensive experiments on machine translation and counter-intuitive arithmetic reasoning tasks demonstrate that MADT consistently outperforms Self-Reflection and Multi-Agent Debate baselines, and that reasoning preservation improves refinement quality and reasoning robustness, validating the effectiveness of synergistic multi-perspective reasoning for complex logical tasks.

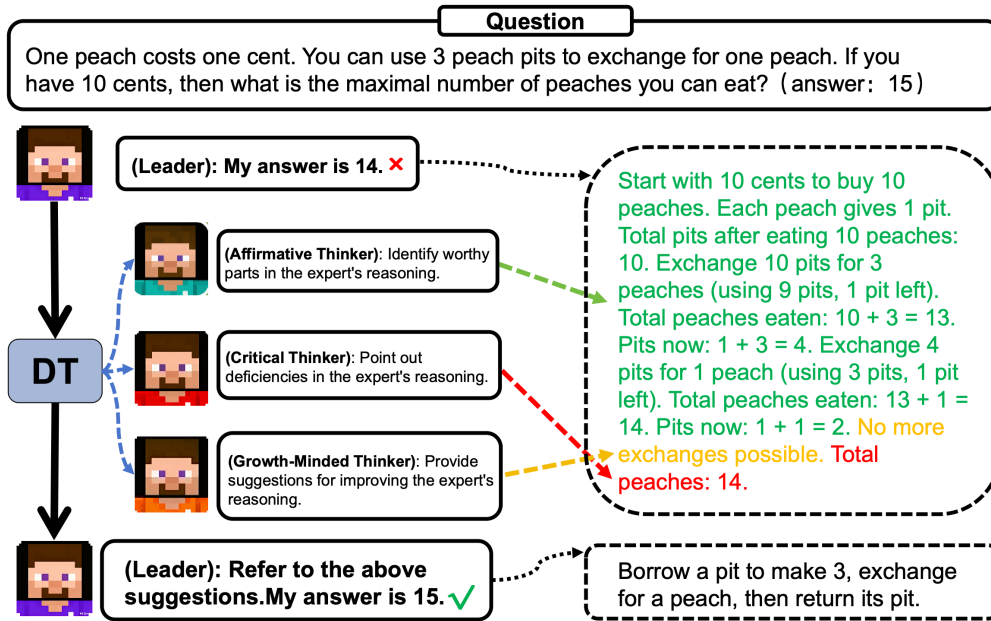


Figure 2: Framework of Multi-Agent Diverse Thinking. We designate  as the Leader,  as the Affirmative Thinker,  as the Critical Thinker, and  as the Growth-Minded Thinker. The content within the dashed box represents the reasoning process of the LLM. The text in green indicates the parts that proponents of the Affirmative Thinker believe are worth retaining. The text in red points out the issues identified by proponents of the Critical Thinker. The text in yellow denotes the parts that proponents of Growth-Minded Thinker believe can be improved.

2 Related Work

Thinking Pattern. The thinking process of LLMs can be regarded as a type of thinking pattern (Li et al., 2023). Thinking Pattern refers to the relatively stable cognitive path and behavioral tendency that individuals or groups maintain when processing information and solving problems (Zhang et al., 2024). Currently, the Self-Reflection pattern representing individual thinking and the Debate pattern representing group cognitive interaction have achieved good results. In the reflect pattern, based on the cognitive closed loop of "retrospection-analysis-improvement", the actor reviews their previous responses, learns lessons from their experience, analyzes the shortcomings of the answer, and improves their answer accordingly. In the Debate pattern, a debate scenario is set up, centering on the process of "view generation - exchange - debate - consensus". First, debaters put forward ideas, then the pro and con sides exchange opinions and start debating, and finally reach a consensus. This paper analyzes the shortcomings of the two existing methods and proposes Diverse Thinking (DT).

Multi-agent System. In human sociology, groups effectively organized by individuals often possess the ability to collaboratively tackle large-scale tasks and accomplish complex objectives with higher

efficiency and better performance than individuals (Chen et al., 2025; Zhou et al., 2024). In the field of artificial intelligence (AI), researchers study the collaboration mechanisms of human societies, with the aim of maximizing the capabilities of individual agents. The division of labor in current multi-agent systems mostly remains at the level of "task assignment" and fails to decompose the agents' "thinking processes" into implementable fine-grained modules (Wu et al., 2023; Zhou et al., 2024). This paper conducts an exploration of the thinking patterns underlying diverse thinking and develops the MADT framework based on this concept. The framework encompasses three different thinkers: affirmative thinking, critical thinking, and growth-minded thinking.

3 Methodology

Figure 2 illustrates the framework of MADT, which consists of a Leader, an Affirmative Thinker, a Critical Thinker, and a Growth-Minded Thinker. Specifically, the Leader proposes a solution to the problem, and each thinker provides thoughts consistent with their own thinking style based on the Leader's solution. The Leader then synthesizes these thoughts and presents the final solution.

Meta Prompts. We use Meta Prompts (denoted as

M) to clarify the task, as well as the participants' thinking patterns and roles in the task. In formulating the prompts, we adopt the structure of "role positioning + task content + execution standards", formalization of which is presented in Equation 1.

$$M = \text{Role Positioning} \oplus \text{Task Content} \oplus \text{Execution Standards} \quad (1)$$

where $\oplus$ denotes the prompt concatenation symbol. Examples of prompts refer to Appendix F. An example of the meta-prompt for the Affirmative Thinker is :

As an Affirmative Thinker proficient in <task>, confirm correct parts of <Leader's answer>. The evaluation criteria are as follows <evaluation criteria>.

Thinker. In our framework, three thinkers with distinct cognitive patterns evaluate the Leader-generated baseline answer comprehensively from multi-dimensional perspectives: a Critical Thinker identifies errors in Large Language Model (LLM) outputs, an Affirmative Thinker preserves valid ideas, and a Growth-Minded Thinker formulates improvement suggestions. Examples of prompts for the affirmative, critical, and Growth-Minded Thinkers are provided in sequence below.

Identify worthy parts in the expert's reasoning.

Point out deficiencies in the expert's reasoning.

Provide suggestions for improving the expert's reasoning.

Leader. A Leader is designed to summarize the suggestions from thinkers and provide the final result. The execution process is shown in Equation 2 and 3.

$$R_n = L(R_{\text{base}}, T_n, M_n) \quad (n = 1, 2, 3) \quad (2)$$

$$R_{\text{final}} = L(R_n, T_{\text{leader}}, M_{\text{leader}}) \quad (3)$$

R_{base} denotes the answer obtained by directly asking the model; T_n stands for role definitions, where $(n = 1, 2, 3)$ represent 3 role definitions with thinking differences; L represents the core execution tool (large language model), which receives input information and role definitions and generates corresponding results; R_n denotes the results generated by the LLM under the corresponding roles; and R_{final} is the final result.

Collaborative Strategy. Multi-agent systems require well-designed collaboration strategies for reasoning frameworks; inappropriate decision-making

will trigger agent conflicts, cause task chain disruption, and ultimately lead to reasoning failure (Wang et al., 2025). The Debate mechanism inherently suffers from the limitation of adversarial viewpoints, with its final output being merely a one-sided solution favoring one party. In contrast, the MADT multi-agent collaboration framework leverages differentiated role division to integrate the strengths of multiple parties for hybrid solution generation, thus achieving a paradigm shift from win-lose confrontation to global optimization. Within MADT, the Leader, as the ultimate decision-maker, executes the collaboration strategy: it retains validated rational ideas based on the inputs of the Affirmative Thinker, acquires impetus for change from the recommendations of the Critical Thinker, and resolves conflicts between the two based on the coordination proposals of the Growth-Minded Thinker.

4 Experiment

4.1 Dataset

We conducted experiments on two challenging tasks—Common MT (He et al., 2020) and CIAR (Liang et al., 2024)—selected for their demand for LLMs to perform in-depth, multi-faceted reasoning, which aligns with our goal of evaluating MADT's ability to enhance complex cognitive processes. For detailed information, please refer to Appendix A.

Commonsense Machine Translation The Common MT dataset focuses on Chinese-to-English translation, with 200 lexical ambiguity examples, 450 contextless syntactic ambiguity examples, and 350 contextual syntactic ambiguity examples (total 1,000 cases). It classifies translations by ambiguity type: lexical ambiguity (driven by single-word polysemy), contextless syntactic ambiguity (structure-based parsing ambiguity without context), and contextual syntactic ambiguity (structure ambiguity activated by context).

Counter-Intuitive Arithmetic Reasoning The CIAR dataset contains 200 questions (each with annotated answers and reasoning processes) and poses two core challenges: (1) resisting intuitive traps (hidden pitfalls test avoidance of superficial, incorrect responses); (2) multi-step reasoning (rigorous step-by-step logic required for correct answers), enabling evaluation of LLMs in complex problem-solving.

Model	Lexical		Contextless		Contextual	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT
Qwen3-32B	83.0	71.1	84.9	74.8	86.0	73.9
Qwen3-30B-A3B	80.1	68.0	83.8	73.1	85.1	73.2
+ Rerank	80.7	68.4	84.7	73.0	85.0	74.1
+ MAPS	82.1	69.9	84.4	73.3	85.1	73.8
+ Self-Reflection	81.2	68.9	83.9	72.4	84.7	73.7
+ MAD	82.0	70.9	84.6	73.7	85.3	74.0
+D-MAD	82.1	71.1	84.8	73.9	84.9	74.2
+MADT	82.2	71.3	84.9	73.9	85.2	74.3
DeepSeek-R1-7B	74.9	62.0	78.3	64.6	80.2	68.2
+ MAD	75.6	62.6	78.6	66.0	81.8	69.9
+D-MAD	75.8	62.7	78.7	66.5	82.0	70.1
+MADT	75.9	63.0	78.9	66.5	82.1	70.3

Table 1: Model Scores of COMET and BLEURT Across Dimensions

4.2 Setting

Input Format. In the parameter configuration of LLMs, the temperature parameter is primarily used to adjust the randomness and creativity of the generated text (Renze, 2024). To ensure the model maintains effective reasoning capabilities while guaranteeing the accuracy and consistency of its outputs, we set the temperature value of LLMs to 0.3.

Evaluation Metrics. For the Common MT dataset, COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) are used as the evaluation metrics, which are widely adopted in LLM translation literature. For the CIAR dataset, ACC is used as the evaluation metric.

Backbone Models. In this work, we primarily utilize four agents within the MADT framework, including one Leader and three thinkers (affirmative, critical, and growth-minded). Three API-based LLMs (Qwen3-32B, Qwen3-30B-A3B and DeepSeek-R1-7B) are evaluated.

4.3 Baselines

We compared the MADT framework with baseline models and also contrasted it with Self-Reflection and Debate methods. Table 1 shows the overall results on the Common MT dataset:

- **Self-Reflection:** This method is a reasoning-enhancement method that enables models to iteratively revise their inference paths by self-evaluating the rationality of intermediate conclusions.
- **Rerank (He et al., 2024):** This method samples LLMs’ translation outputs four times and

selects the optimal candidate based on the quality estimation metric HUMANr5.

- **MAPS (He et al., 2024):** This method empowers LLMs to simulate the human translation workflow by performing pre-translation analysis, a paradigm analogous to applying chain-of-thought reasoning to the translation task.
- **MAD:** This method establishes a debate scenario where debaters first propose ideas, followed by an exchange of opinions between the pro and con sides to initiate the debate, and finally a consensus is reached.
- **D-MAD:** Within the MAD framework, each participating agent embodies distinct thinking through different reasoning methods.

4.4 Result

Results on Common MT. The experimental results demonstrate that MADT exhibits remarkable generality and stability: On the high-performance Qwen3-30B-A3B model, compared with the original model, it achieves a significant improvement of 2.1 points (80.1→82.2) in the COMET metric and 3.3 points (68.0→71.3) in the BLEURT metric in the Lexical dimension. On the mid-low performance DeepSeek-R1-7B model, MADT also maintains stable positive optimization. In the Contextual dimension, it increases the COMET and BLEURT metrics by 1.9 points (80.2→82.1) and 2.1 points (68.2→70.3) respectively, with a relative improvement rate significantly higher than that of

high-performance models, reflecting its optimization potential for weak base models. A further comparison of the effects of MAD and MADT shows that the optimization effect is most prominent, especially in the lexical level.

Results on CIAR. To evaluate the performance of different methods in handling Counter-Intuitive Arithmetic Reasoning (CIAR), we conducted systematic experiments on the CIAR dataset and report quantitative accuracy results. As shown in Table 2, among all tested methods, Qwen3-32B achieves a competitive accuracy of 68.0%. The baseline model, DeepSeek-R1-7B, obtains a baseline accuracy of 50.0%. When the baseline is augmented with individual modules, the results are as follows: combining it with Self-Reflection increases the accuracy to 56.0%; combining it with MAD boosts the accuracy to 61.0%; and combining it with D-MAD elevates the accuracy to 64.0%. Additionally, the baseline combined with MADT achieves a notable accuracy of 66.0%, which demonstrates the latter’s superior effectiveness in enhancing CIAR capability.

Method	Acc%
Qwen3-32B	68.0
DeepSeek-R1-7B	50.0
+ Self-Reflection	56.0
+ MAD	61.0
+D-MAD	64.0
+MADT	66.0

Table 2: Accuracy Comparison of Different Methods on CIAR Dataset

4.5 Ablation Study

To verify the effectiveness of each core component in the proposed MADT framework, we conducted systematic ablation experiments on the target task. Taking DeepSeek-R1-7B as the baseline model, we analyzed the independent contribution of each module to the model performance by gradually removing the key modules of MADT. The experimental results are summarized in Table 3.

Importance of the Critical Thinker. When this module is removed, the model’s accuracy drops sharply from 66.0% to 53.0%, a decrease of 13.0 percentage points. The COMET and BLEURT scores also decline significantly. This data indicates that the Critical Thinker is a crucial pillar of the model’s capabilities.

Method	Acc	COMET _{avg}	BLEURT _{avg}
DeepSeek-R1-7B	50.0	77.80	64.93
+MADT	66.0	78.97	66.60
w/o Affirmative	62.0	78.52	66.31
w/o Critical	53.0	78.13	65.62
w/o Growth-Minded	60.0	78.81	66.13

Table 3: Performance Comparison of Different Methods (Acc, COMET, BLEURT)

Importance of the Growth-Minded Thinker. After ablating this module, the model’s accuracy decreases from 66.0% to 60.0%, a drop of 6.0 percentage points. Additionally, the COMET and BLEURT scores show varying degrees of regression. This illustrates that the Growth-Minded Thinker incorporates optimization strategies, which can effectively assist the model in achieving adaptive performance improvement in complex task scenarios.

Importance of the Affirmative Thinker. Notably, although the Affirmative module in this framework only focuses on analyzing the parts of the answer that are worth retaining, its existence is still indispensable. Data show that removing the Affirmative module leads to a decrease in the model’s accuracy from 66.0% to 62.0%, a drop of 4.0 percentage points, with corresponding declines in the COMET and BLEURT scores. This indicates that, by enhancing the identification and retention of correct output components, the Affirmative module provides important support for the stability of model performance and serves as an indispensable part of the MADT framework.

5 Analysis

5.1 Diversity Analysis

LLMs can fall into a self-cognition deadlock and become reluctant to revise their answers. Opposing viewpoints in debates may also constrain responses and inhibit the possibility of generating alternative solutions (Yin et al., 2020). Translation diversity can be evaluated using the Self-BLEU score. Its principle is to calculate the n-gram overlap between generated samples to reflect the diversity of text generation models; a lower score indicates greater differences between samples. We calculate the Self-BLEU values of initial translations and final translations, and then derive the diversity score. The specific formulation is given in Equation 4, as

follows:

$$Diversity = 100 - Self_BLEU(R_{base}, R_{final}) \times 100 \quad (4)$$

where *Diversity* denotes the final diversity score, and *Self_BLEU* corresponds to the metric adopted in our method.

Method	Diversity
Self-Reflection	25.8
MAD	30.8
MADT	47.1

Table 4: Diversity Comparison of Different Methods

As shown in Table 4, compared with Self-Reflection and MAD, our proposed MADT more effectively addresses limitations in reasoning (cognitive degradation and opposing perspectives) and enables substantial improvements in answers.

5.2 The Driving Force for Change

The Critical Thinker serves as the primary force driving performance improvements within the MADT framework. As shown in Table 5, by engaging in critical evaluation of the results, it identifies errors and rectifies them. In this case, for example, the core expression conveys the semantic requirement of “completely annihilating an enemy combat unit.” Although the translation “Eat up an enemy division” generated by Qwen3-32B achieves literal correspondence, it exhibits a significant semantic deviation. The inclusion of the Critical Thinker corrects this core error in the original translation.

Source/Method	Translation
Source	吃掉敌人一个师。
Correct Ref.	Destroy a division of the enemy.
Incorrect Ref.	Eat a division of the enemy.
Qwen3-32B	Eat up an enemy division.
+Critical	Destroy an enemy division.
+Growth-Minded	Defeat an enemy division.
+MADT	Destroy an enemy division.

Table 5: Translation Comparison of Different Modules. Green font: correctly translated tokens; red font: mis-translated tokens. Background colors indicate thinkers’ judgements: green (Affirmative Thinker: correct), red (Critical Thinker: erroneous), yellow (Growth-Minded Thinker: room for improvement). This scheme holds for following examples.

5.3 Preserve Correct Thinking.

Within the MADT framework, the Critical Thinker may sometimes take an opposing perspective by

affirming the presence of errors, and the doubts raised may not always be entirely correct. **The role of the Affirmative Thinker therefore proves crucial: rather than simply defending the original answer, it evaluates and preserves its rational components.** As shown in Table 8, when the Critical Thinker proposes “go into the water,” this translation weakens the key information of “swimming.” The inclusion of the Affirmative Thinker not only prevents potential semantic deviations caused by critical thinking but also enhances the accuracy and practicality of the final answer through collaboration with the Growth-Minded Thinker.

5.4 Refinement Mechanism

The Growth-Minded Thinker focus more on the details that require improvement. While Critical Thinker verifies the absence of semantic errors, those with a growth mindset acutely identify subtle opportunities for optimization between surface-level expressions and underlying needs. As shown in Table 9, the colloquial idiom “keeps it hot” can convey the meaning of “maintaining momentum,” but it lacks precision and professionalism in formal contexts. Replacing the vague phrase “keeps it hot” with the more precise term “maintains high intensity” enhances the formality of the expression and its suitability for the given context.

5.5 Comparison with Self-Reflection

To address judgment biases arising from the Self-Reflection mechanism’s reliance on a single cognitive model, we introduce external content to expand cognitive boundaries. Unlike Self-Reflection, our method integrates three-dimensional cognitive perspectives (affirmative, critical, Growth-Minded Thinkers): affirming core rational components to reinforce strengths, identifying errors, and proposing feasible improvements, forming a “recognition-correction-enhancement” pathway that mitigates standalone Self-Reflection’s cognitive blind spots. We quantify agent performance with two metrics: Self-BLEU (cognitive diversity between an agent’s translations and final output) and COMET score difference (direct contributions vs. initial translations). For the Affirmative Thinker—which provides only affirmative judgments rather than specific answers—direct contribution is computed via ablation experiments, while Self-BLEU is calculated by comparing final results before and after ablation.

As shown in Figure 3, under the Self-Reflection

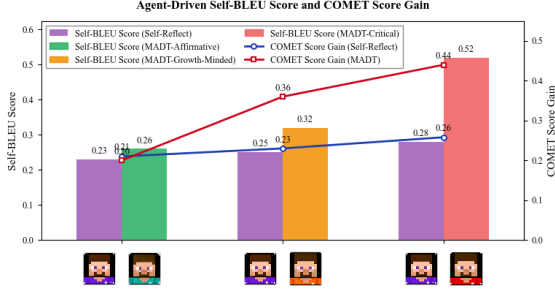


Figure 3: The gains of different agents on COMET scores and Self-BLEU scores in the Self-Reflection method and our method

mechanism, the maximum COMET score gain is only 0.26, and the Self-BLEU score maintains an interval of 0.23 - 0.28, indicating that the agent’s improvement on the result is relatively limited. In contrast, in our method (MADT), the Critical Thinker achieves a COMET score gain as high as 0.44 and a Self-BLEU score of 0.52, while the Growth-Minded Thinker achieves a COMET score gain of 0.36 and a Self-BLEU score of 0.32. By leveraging multi-faceted thinking dimensions, it realizes a distinctive and significant enhancement, fully demonstrating the advantage of the multi-cognitive architecture in performance gain.

5.6 Comparison with Debate

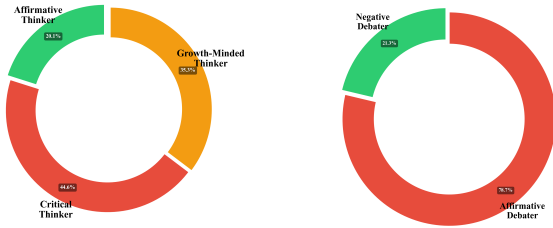


Figure 4: Proportion of Each Agent’s Influence on the Final Result (Left: Our Method; Right: Debate Method)

The adversarial nature of debates is repeatedly confined to just two lines of thinking. Moreover, incorrect viewpoints may reject or suppress lines of reasoning proposed by correct ones. Compared with the MAD method, our framework is not limited to mere affirmation or negation of viewpoints; instead, it conducts fine-grained decomposition and multi-dimensional deepening of the original answer. By examining what is correct, what is incorrect, and what is worthy of improvement, it avoids the limitations of adversarial thinking and brings the answer

closer to perfection. To quantify the exclusive value of each agent to the final result, we innovatively designed the Unique Agent Contribution indicator(UAC). This metric is calculated by subtracting the Self-BLEU score between the final output and the initial translation from the Self-BLEU score between the output of a specific agent and the final result. The specific equations are shown in Equations 5 and 6.

$$Base = Self_BLEU(R_{base}, R_{final}) \quad (5)$$

$$UAC = Self_BLEU(R_n, R_{final}) - Base \quad (6)$$

The final results are presented in Figure 4. In MAD, the agents’ influence on the final outcome exhibits a distinct binary confrontational characteristic. The proportional discrepancy between affirmative debaters (78.7%) and negative debaters (21.3%) reflects that the outcome is constrained by the confrontational thinking of mere affirmation or negation of viewpoints, which tends to fall into an either-or mode of thinking. In contrast, our method integrates multiple thinking dimensions such as critical, growth-minded, and affirmative thinking. The more diverse distribution of Critical Thinkers (44.6%), Growth-Minded Thinkers (35.3%), and Affirmative Thinkers (20.1%) indicates that, our method, through diverse thinking, breaks free from the thinking framework of mere affirmation or negation and overcomes the limitations of confrontational reasoning.

6 Conclusion

In this paper, we address two critical limitations of existing LLM cognitive paradigms: the cognitive degradation and single-mode bias of Self-Reflection mechanisms, and the binary adversarial constraints of Multi-Agent Debate (MAD) frameworks that suppress diverse perspectives. To resolve these issues, we propose the Multi-Agent Diverse Thinking (MADT) framework, which innovatively integrates three specialized thinking roles—Affirmative Thinker, Critical Thinker, and Growth-Minded Thinker to establish a "recognition-correction-enhancement" cognitive loop. This design transcends the either/or limitation of traditional debate methods and the self-cognitive deadlock of single-agent reflection, unlocking more possibilities for multi-agent collaborative optimization.

Limitations

Building upon sufficient thinking time and computing resources for LLMs, this work enables the agent to derive the final answer through multiple rounds of interaction. Within the framework, the Leader is tasked with integrating analyses from various thinkers to formulate conclusions. Notably, long question input remains a persistent challenge for LLMs. To mitigate this issue, this paper introduces a prompt design that constrains thinkers to convey only the core reasoning processes and essential answers, thereby effectively alleviating the aforementioned challenge.

Acknowledgments

The work is supported by grants from National Natural Science Foundation of China (No. 62372189).

References

- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.
- Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2025. [Multi-agent consensus seeking via large language models](#). *Preprint*, arXiv:2310.20151.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. [Chain-of-verification reduces hallucination in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujia Yang, Nan Duan, and Weizhu Chen. 2023. [CRITIC: Large language models can self-correct with tool-interactive critiquing](#). In *Second Agent Learning in Open-Endedness Workshop*.
- Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. [The box is in the pen: Evaluating commonsense reasoning in neural machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujia Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. [Exploring human-like translation strategy with large language models](#). *Transactions of the Association for Computational Linguistics*, 12:229–246.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *CoRR*, abs/2009.03300.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. [CAMEL: Communicative agents for “mind” exploration of large language model society](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujia Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.
- Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025. [Breaking mental set to improve reasoning through diverse multi-agent debate](#). In *The Thirteenth International Conference on Learning Representations*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Matthew Renze. 2024. [The effect of sampling temperature on problem solving in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). *Preprint*, arXiv:2210.09261.
- Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D. Nguyen. 2025. [Multi-agent collaboration mechanisms: A survey of llms](#). *Preprint*, arXiv:2501.06322.
- Haochun Wang, Sendong Zhao, Jingbo Wang, Zewen Qiang, Bing Qin, and Ting Liu. 2025. [Beyond frameworks: Unpacking collaboration strategies in multi-agent systems](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21361–21375, Vienna, Austria. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. [Autogen: Enabling next-gen llm applications via multi-agent conversation](#). *Preprint*, arXiv:2308.08155.
- Haiyan Yin, Dingcheng Li, Xu Li, and Ping Li. 2020. [Meta-cotgan: A meta cooperative training paradigm for improving adversarial text generation](#). *Preprint*, arXiv:2003.11530.
- Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. [Exploring collaboration mechanisms for LLM agents: A social psychology view](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.
- Yuyan Zhou, Liang Song, Bingning Wang, and Weipeng Chen. 2024. [MetaGPT: Merging large language models using model exclusive task arithmetic](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1724, Miami, Florida, USA. Association for Computational Linguistics.
- Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaying Zhang, and Yujia Yang. 2023. [Solving math word problems via cooperative reasoning induced language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4471–4485, Toronto, Canada. Association for Computational Linguistics.

A Dataset

We conducted experiments on two challenging tasks, including Common MT and CIAR, which require LLMs to engage in in-depth reasoning.

A.1 Common MT

The Common MT Dataset is a Chinese-to-English translation dataset, which provides Chinese-to-English translation cases categorized into three types based on translation ambiguity properties: lexical ambiguity, contextless syntactic ambiguity, and contextual syntactic ambiguity. Specifically, the dataset contains 200 lexical ambiguity examples, 450 contextless syntactic ambiguity, and 350 contextual syntactic ambiguity examples.

Lexical Ambiguity. This refers to the phenomenon where a sentence yields multiple interpretations due to a single word possessing several mutually independent core meanings. The root cause of such ambiguity lies solely at the lexical level, having no connection with the sentence structure.

contextless syntactic ambiguity. This denotes the phenomenon where a sentence generates multiple interpretations because its syntactic structure (e.g., constituent segmentation, grammatical relations) allows for multiple valid parsing methods. The ambiguity originates exclusively at the syntactic level and can be identified without relying on contextual information.

contextual syntactic ambiguity. This describes the phenomenon where a sentence’s syntactic structure inherently has multiple potential parsing possibilities; however, the ambiguity is temporarily concealed when separated from specific contexts and becomes activated once concrete contextual information is introduced.

A.2 CIAR

The CIAR dataset comprises 200 questions, each paired with a corresponding answer and detailed reasoning process, and it presents two core challenges for models: first, resisting intuitive pitfalls, where questions are crafted with hidden traps to test whether LLMs can resist the temptation of superficial cues and avoid selecting intuitively appealing yet incorrect answers; second, multi-step reasoning capability, as each correct answer requires a rigorous multi-step reasoning process, thereby enabling the evaluation of LLMs’ performance in complex decision-making and problem-solving scenarios.

B Experimental Results

This appendix presents additional experimental results and case analyses referenced in the main text, including Tables 8 and 9.

B.1 Human Evaluation on Common MT Benchmark

To mitigate the inherent limitations of automatic metrics for ambiguous translation tasks, we conduct human-evaluation experiments following the evaluation protocol from the MAD paper. We randomly shuffled translation outputs generated by all compared methods and recruited three professional translators to provide independent ratings. The inter-annotator agreement (Krippendorff’s Alpha) reaches 0.72. Final scores are computed as the average of annotator ratings. We evaluate translation quality from three dimensions: *Lexical*, *Contextless*, and *Contextual*. All human scores range from 1 to 5, with higher values indicating better translation quality. Detailed human evaluation results are reported in Table 6.

Method	Lexical	Contextless	Contextual
Qwen3-32B	3.32	3.41	3.39
+Rerank	3.12	3.21	3.18
+MAPS	3.18	3.23	3.20
+Self-Reflection	3.22	3.28	3.25
+MAD	3.35	3.44	3.41
+D-MAD	3.37	3.48	3.41
+MADT	3.51	3.53	3.53
DeepSeek-R1-7B	2.83	2.88	2.91
+MAD	2.90	2.96	2.93
+D-MAD	2.91	2.98	3.01
+MADT	2.98	3.04	3.12

Table 6: Human evaluation scores (1–5, higher is better) on Common MT benchmark.

The human evaluation results further demonstrate that our MADT consistently outperforms all baseline methods across all evaluation dimensions for both backbone models, yielding improved human-perceived translation quality.

B.2 Partial-Correct Reasoning Preservation Analysis

To provide direct evidence for our core mechanism that valid partially-correct reasoning steps can be preserved during collaborative refinement, we conduct a fine-grained step-level analysis on the CIAR benchmark. We randomly sample reasoning cases and decompose the initial reasoning outputs into individual fine-grained steps. Each reasoning step is independently annotated as correct or incorrect by three researchers specialized in LLM reasoning, with annotation conflicts resolved via majority voting. Three evaluation metrics are defined for comparison:

- **Preservation Rate (PR)** ($\uparrow$): The percentage of initially-correct reasoning steps that remain correct after collaborative refinement.
- **Corruption Rate (CR)** ($\downarrow$): The percentage of initially-correct reasoning steps that are incorrectly modified during collaborative refinement.
- **Improvement Rate (IR)** ($\uparrow$): The percentage of initially-incorrect reasoning steps that are successfully corrected.

Method	Preservation Rate $\uparrow$	Corruption Rate $\downarrow$	Improvement Rate $\uparrow$
MAD	46.3	13.6	23.1
D-MAD	45.1	17.2	29.7
MADT (Ours)	53.5	9.8	25.3

Table 7: Step-level reasoning preservation analysis on CIAR.

As shown in Table 7, MADT achieves the highest preservation rate and the lowest corruption rate among all compared methods. This demonstrates that the Affirmative Thinker component can effectively retain valid reasoning segments and prevent them from being damaged in the refinement loop. Meanwhile, MADT maintains a competitive improvement rate, which indicates our framework is capable of fixing flawed reasoning while preserving correct steps. This step-level analysis offers direct empirical validation for our proposed collaborative refinement mechanism.

Source	这几个学生约好一起去水库下水。
Correct Ref.	these students agreed to go to the reservoir together for swimming.
Incorrect Ref.	these students agreed to go to the reservoir to do bad things.
Qwen3-32B	these students agreed to go together to the reservoir to go swimming.
+Critical	some students agreed to go to the reservoir together to go into the water.
+Growth-Minded	these students agreed to go to the reservoir together to swim.
+MADT	some students agreed to go to the reservoir together to go swimming.

Table 8: Translation Comparison of Different Modules

Source	他对这件事情的热度持续很久。
Correct Ref.	His enthusiasm for this matter lasted for a long time.
Incorrect Ref.	His temperature for this matter lasted for a long time.
Qwen3-32B	He keeps it hot for a long time.
+Critical	He keeps it hot for a long time.
+Growth-Minded	He maintains high intensity for a long time.
+MADT	He maintains high intensity for a long time.

Table 9: Translation Comparison of Different Methods

C Extra Computational Cost

We quantify the reasoning cost by the number of generated tokens. As shown in Table 10, Self-Reflection incurs a cost that is 1.83 times that of the baseline Chain-of-Thought (CoT) method, while the cost of the MADT method reaches 2.72 times that of the baseline CoT. Given the considerable optimization potential of our proposed method, this increase in cost is deemed acceptable.

MADT consumes $2.72\times$ the tokens of standard CoT, which is higher than MAD ($2.46\times$) and Self-Reflection ($1.83\times$). A critical question is whether MADT’s performance gains arise from its collaborative structure or merely from its larger computational budget. To answer this, we conduct a conservative controlled experiment: as shown in Table 11, we increase the token budgets of MAD and Self-Reflection to match or exceed MADT’s budget, and then re-evaluate their performance. If MADT still outperforms these baselines under equal or higher budgets, its structural advantage is confirmed.

Method	Cost (Tokens)
CoT	$1.0\times$
Self-Reflection	$1.83\times$
MAD	$2.46\times$
MADT	$2.72\times$

Table 10: Comparison of method costs.

Method	Acc	COMET _{avg}	BLEURT _{avg}
DeepSeek-R1-7B	50.0	77.80	64.93
+Rerank	58.0	77.12	65.21
+MAPS	60.0	77.35	65.82
+Self-Reflection	60.0	77.31	65.67
+MAD	63.0	78.66	66.17
+D-MAD	64.0	78.73	66.31
+MADT(Role-shuffled)	66.0	78.93	66.58
+MADT (Ours, Equal Budget)	66.0	78.97	66.60

Table 11: Cost-controlled performance comparison: baselines are given equal or higher token budgets than MADT

D Latency Analysis

To evaluate the practical deployability of MADT, we measure the average inference latency (wall-clock time) on the CIAR dataset using a single RTX 3090 GPU with DeepSeek-R1-7B. The results are shown in Table 12. All methods are implemented under the same API calling infrastructure; latency is averaged over 50 randomly sampled questions.

Method	Latency ($\times$ CoT)
CoT	$1.00\times$
Self-Reflection	$1.53\times$
MAD	$2.50\times$
MADT	$1.86\times$

Table 12: Comparison of average inference latency.

A key architectural difference affects latency: **MAD** requires sequential debate rounds (e.g., Agent A $\rightarrow$ Agent B $\rightarrow$ Agent A $\dots$), where each turn must wait for the previous agent’s re-

Metric	CoT	Self-Reflection	MAD	MADT
Accuracy (%)	76.8 $\pm$ 0.5	79.0 $\pm$ 0.5	81.3 $\pm$ 0.5	82.1 $\pm$ 0.5

Table 13: Comparison of accuracy on the “abstract_algebra” subset of MMLU.

Method	Symbolic Reasoning (BBH)		
	Penguin	Date	Colored Objects
CoT	60.3	58.4	58.2
Self-Reflection	62.3	59.3	59.1
MAD	63.7	64.2	59.5
MADT	64.3	65.8	59.6

Table 14: Comparison of accuracy on symbolic reasoning tasks from BBH.

Metric	CoT	Self-Reflection	MAD	MADT
Accuracy (%)	72.7 $\pm$ 0.5	74.1 $\pm$ 0.5	75.6 $\pm$ 0.5	76.2 $\pm$ 0.5

Table 15: Comparison of accuracy on GSM8K.

sponse, resulting in high wall-clock time. In contrast, **MADT**’s three thinkers (Affirmative, Critical, and Growth-Minded) operate independently on the same initial answer and can be executed **in parallel**, with only the Leader needing a final sequential aggregation. As shown in Table 12, the serial implementation of MADT (thinkers executed one after another) already achieves a latency of only $1.86\times$ CoT, which is substantially lower than MAD’s $2.50\times$ CoT. This counter-intuitive result stems from two factors: (1) MAD’s debate rounds involve full-length responses from multiple agents, whereas each MADT thinker focuses on a specialized subtask and thus generates shorter outputs; (2) when deployed with concurrency, MADT not only improves reasoning accuracy but also offers favorable latency characteristics compared to traditional debate frameworks.

E Expanded Experiments

E.1 Results on the subset “abstract_algebra” of MMLU

To further verify the effectiveness of MADT in complex reasoning scenarios, we conduct additional experiments on challenging mathematical tasks. MMLU (Massive Multitask Language Understanding) is a comprehensive multi-disciplinary benchmark covering 57 subjects across STEM, humanities, social sciences, and professional domains, which is widely adopted to evaluate the general

knowledge and advanced reasoning capabilities of large language models (Hendrycks et al., 2020).

We select the *abstract algebra* subset from MMLU, a high-difficulty mathematical domain containing college-level problems, and adopt Qwen3-32B as the backbone model. All experimental results are averaged over three repeated runs to report the mean accuracy and standard deviation for a fair comparison. The detailed results are presented in Table 13.

E.2 Results on symbolic reasoning tasks

To further verify the generalization and robustness of MADT on complex symbolic and logical reasoning, we conduct additional experiments on three challenging subsets from BIG-Bench Hard (BBH) (Suzgun et al., 2022). BBH is a widely recognized tough benchmark that focuses on multi-step deduction, state tracking, and structured reasoning tasks where large language models perform significantly worse than humans. We select three representative and cognitively challenging tasks: *Penguins in a Table*, *Date Understanding*, and *Reasoning about Colored Objects*. These tasks involve table-based reasoning, temporal logic inference, and dynamic object state tracking, respectively, which require rigorous multi-hop thinking and are highly suitable for evaluating the synergized reasoning ability of our framework.

We adopt DeepSeek-R1-7B as the backbone

model and perform all evaluations under the consistent few-shot setting. All experimental results are averaged over three repeated runs to report the mean accuracy and standard deviation for a fair comparison. The detailed results are presented in Table 14.

E.3 Results on math

To further verify the effectiveness of MADT in mathematical reasoning scenarios, we conduct additional experiments on typical multi-step arithmetic reasoning tasks. GSM8K (Grade School Math 8K) is a standard benchmark for evaluating multi-step mathematical reasoning ability, which consists of 8,500 high-quality elementary school math word problems. It requires 2–8 steps of continuous deductive reasoning and only involves four fundamental operations, making it widely adopted to test the logical reasoning and numerical calculation capabilities of large language models (Cobbe et al., 2021).

We select the GSM8K dataset to validate the rationality and advancement of our framework, and adopt DeepSeek-R1-7B as the backbone model. All experimental results are averaged over three repeated runs to report the mean accuracy and standard deviation for a fair comparison. The detailed results are presented in Table 15.

E.4 Results on Larger Backbone Model

To validate the scalability and universal effectiveness of our MADT framework, we further conduct supplementary experiments on a substantially stronger backbone model, DeepSeek-R1-70B. Previous evaluations are mainly performed on medium-sized models, which leaves the scalability of our method on large-scale models unexplored. We conduct experiments on mathematical reasoning and machine translation benchmarks under strictly consistent settings. All results are averaged over three independent runs for fair and stable comparison. The detailed results are presented in Table 16.

Method	Acc	COMET_avg	BLEURT_avg
DeepSeek-R1-70B	80.00	86.32	76.45
DeepSeek-R1-70B+MADT	83.00	87.19	78.21

Table 16: Main results on DeepSeek-R1-70B

As observed, our MADT method yields consistent performance improvements across all metrics. This confirms that the proposed collaborative refinement mechanism is not limited to small

or medium-sized backbones, but generalizes well to stronger large-scale reasoning models, further verifying the robustness and applicability of our framework.

F Meta Prompt

In formulating the prompts, we adopt the structure of "role positioning + task content + execution standards"

F.1 Meta Prompt for Common MT

Leader’s meta-prompt:

As a Chinese-English translation expert familiar with <task>’s text type, find the correct English translation of <task> (balance accuracy, fluency, context).

Affirmative Thinker’s meta-prompt:

As an Affirmative Thinker proficient in English and <task>’s text type, confirm correct parts of <task>’s English translation. Evaluate : 1.Accuracy (full meaning, no additions/omissions); 2.Fluency (English habits/norms). Provide specific examples (e.g., 'XX→YY is accurate').

Critical Thinker’s meta-prompt:

As a Critical Thinker proficient in Chinese-English and cultural differences, identify flaws in <task>’s English translation. Evaluate: 1.Accuracy (mistranslation, missing meaning, poor cultural adaptation); 2.Fluency (awkward expressions). Provide evidence (e.g., 'XX→YY is wrong';).

Growth-Minded Thinker’s meta-prompt:

As a Growth-Minded Thinker, optimize <task>’s English translation. Steps: 1.Recognize strengths (with examples); 2.Propose targeted improvements : - Accuracy (adjust technically correct but mismatched terms); - Fluency (revise understandable but unsmooth expressions). Format: 'Current issue + Revised content + Reason' (e.g., 'Current: ...function of... / Revised: ...features...').

F.2 Meta Prompt for CIAR

Leader's meta-prompt:

As a math problem-solving expert, find the correct answer to <task> (balance logical rigor, calculation accuracy, problem context).

Affirmative Thinker's meta-prompt:

As an Affirmative Thinker proficient in math, confirm correct parts of <task>'s answer. Evaluate: 1. Logical rigor (no logical flaws, correct formula application); 2. Calculation accuracy (no miscalculations, correct unit handling). Provide specific examples (e.g., 'XX step correctly applies the average speed formula').

Critical Thinker's meta-prompt:

As a Critical Thinker proficient in math, identify flaws in <task>'s answer. Evaluate: 1. Logical flaws (e.g., wrong formula selection, ignored conditions); 2. Calculation errors (e.g., miscalculations, incorrect unit conversion). Provide evidence (e.g., 'XX step confuses average speed with average of speeds, leading to wrong results').

Growth-Minded Thinker's meta-prompt:

As a Growth-Minded Thinker, optimize <task>'s answer. Steps: 1. Recognize strengths (with examples, e.g., 'XX step clearly lists total distance and time for average speed'); 2. Propose targeted improvements: - Logical clarity (adjust confusing reasoning steps); - Calculation conciseness (revise redundant calculation processes). Format: 'Current issue + Revised content + Reason' (e.g., 'Current: Complex fraction simplification is unclear / Revised: Simplify $2d/(4d/3)$ to $3/2$ directly / Reason: Reduces redundant calculation steps').

G Thinking Process

G.1 Common MT

The following presents the translation process of MADT for the sentence "放学后这几个学生约

好一起去水库下水". First, clear translation task objectives are established, with the analysis conducted from three dimensions: accuracy, fluency, and contextual adaptability. The Leader first proposes an initial translation. Then, the Affirmative Thinker provides justifications for validating this initial version. Next, the Critical Thinker identifies the limitations of the initial translation and offers an alternative version. Subsequently, the Growth-Minded Thinker puts forward targeted improvement suggestions and presents a revised translation. Finally, the Leader synthesizes the rationales from all thinkers and determines the final translation.

<task>: Translate 放学后这几个学生约好一起去水库下水。 from Chinese to English.

 "reasoning": "1. **Accuracy ensured via term matching and cultural retention**: '约好' is translated as 'agreed to' to preserve the mutual arrangement implied in Chinese. '下水' is rendered as 'to swim' (contextual to reservoirs) instead of a literal 'go into the water', retaining activity specificity. 2. **Fluency via English habits**: The structure 'After school, these students agreed to...' mirrors natural English narrative flow. 'Go together' avoids awkward literalism (e.g., 'went together') by using present infinitive for clarity. 3. **Adaptation to source text type (narrative)**: The translation preserves the casual, descriptive tone of the original Chinese sentence, which recounts a planned group activity. No legal or formal rigor is needed, but vividness is maintained through active verbs ('agreed', 'go').", "translation": "After school, these students agreed to go together to the reservoir to swim."

 "reasoning": "1. Accurate parts: '约好' translated as 'agreed to' captures the mutual arrangement (vs. vague 'planned to'). '下水' rendered as 'to swim' matches cultural context (vs. literal 'go into water'). 2. Fluent parts: 'Agreed to go together' uses common English phrasing for planned group actions (vs. awkward 'had agreed to go together'). 'To the reservoir to swim' follows English prepositional clarity (vs. fragmented 'reservoir and swim'). 3. Text type adaptation strengths: Active verbs ('agreed', 'go') preserve the narrative's casual, action-oriented tone. Simple declarative structure mirrors Chinese's straightforward storytelling style."

 "reasoning": "1. **Accuracy deficiencies**: '下水' is translated as 'to swim' which may omit core meaning if the activity isn't specifically swim-

ming (e.g., wading, boating). '约好' as 'agreed to' assumes prior arrangement but the original implies a group decision, not necessarily a formal agreement. '这几个学生' is translated as 'these students' but the Chinese phrase is more neutral (could mean 'some students'). 2. ****Fluency deficiencies****: 'Go together to the reservoir' is awkward; natural English would use 'go to the reservoir together.' The phrase 'agreed to go together to the reservoir to swim' is verbose and disrupts flow. 3. ****Reasoning deficiencies****: No explanation of why '下水' was narrowed to 'swim' despite potential contextual ambiguity. The reasoning claims 'cultural retention' but fails to address if 'reservoir' is an appropriate term in the target language's cultural context (e.g., 'water reservoir' vs. 'lake'). The claim about 'source text type (narrative)' is vague and lacks evidence of tone adaptation.", "translation": "After school, some students agreed to go to the reservoir together to go into the water."

 "reasoning": "1. ****Recognized strengths****: 'agreed to' accurately captures '约好' by emphasizing mutual planning; 'to swim' retains contextual specificity for '下水' in a reservoir setting. The structure 'After school, these students agreed to...' aligns with English narrative flow. 2. ****Optimization areas****: (a) 'go together to the reservoir' → 'go to the reservoir together' (revised word order for natural English phrasing). (b) 'to swim' → 'to go swimming' (more idiomatic for recreational water activity). 3. ****Reasoning supplementation****: The expert's reasoning omitted the idiomatic adjustment for '下水' and word order fluency. Adjusting to 'go swimming' clarifies the activity as leisure, while reordering 'together' enhances readability.", "translation": "After school, these students agreed to go to the reservoir together to go swimming."

 "reasoning": "1. ****Adopted thinker B's corrections****: Replaced 'these students' with 'some students' to neutralize specificity (B's accuracy point). Adjusted 'go together to the reservoir' to 'go to the reservoir together' for natural word order (B's fluency point). 2. ****Integrated thinker C's optimizations****: Replaced 'to swim' with 'to go swimming' for idiomatic clarity (C's fluency optimization). 3. ****Retained thinker A's strengths****: Preserved 'agreed to' for '约好' (mutual arrangement) and retained 'reservoir' as the specific water body. 4. ****Contextual adaptation****: 'Go swimming' clarifies leisure activity vs. literal 'go into water,' while 'some students' avoids over-specifying the

group. The declarative structure mirrors Chinese's straightforward narrative tone.", "translation": "After school, some students agreed to go to the reservoir together to go swimming."

G.2 CIAR

The following presents the reasoning process of MADT for solving the empty peach pit exchange problem.

<task>: One peach costs one cent. You can use 3 peach pits to exchange for one peach. If you have 10 cents, then what is the maximal number of peaches you can eat?

 "reasoning": "Start with 10 cents to buy 10 peaches. Each peach gives 1 pit. Total pits after eating 10 peaches: 10. Exchange 10 pits for 3 peaches (using 9 pits, 1 pit left). Total peaches eaten: $10 + 3 = 13$. Pits now: $1 + 3 = 4$. Exchange 4 pits for 1 peach (using 3 pits, 1 pit left). Total peaches eaten: $13 + 1 = 14$. Pits now: $1 + 1 = 2$. No more exchanges possible. Total peaches: 14.", "answer": "14"

 "reasoning": "1. Logically rigorous parts: The expert correctly accounts for the iterative exchange process, ensuring that each exchange of pits for peaches is tracked alongside leftover pits. For example, after the first exchange of 10 pits, the reasoning explicitly notes using 9 pits (3 peaches) and retaining 1 pit, avoiding overcounting. 2. Calculationally accurate parts: The step where 10 pits are exchanged for 3 peaches (using 9 pits) and the subsequent addition of 3 new pits to the leftover 1 pit (totaling 4) is mathematically precise. The division of 4 pits into 1 exchange (3 pits) with 1 leftover is also correct. 3. Problem type adaptation strengths: The reasoning adapts to the recursive nature of the problem by repeatedly applying the exchange rule (3 pits → 1 peach) and tracking leftovers, which aligns with the problem's requirements."

 "reasoning": "1. Logically rigorous parts: The reasoning strictly tracks the iterative exchange process of pits for peaches, clearly distinguishing between initial pits, used pits for exchange, leftover pits, and newly generated pits after eating peaches. For example, it explicitly records that 10 initial pits use 9 for the first exchange (obtaining 3 peaches) and retain 1, then updates the total pits to 4 by adding 3 new pits from eaten peaches, avoiding logical omissions or confusion. It also considers

two scenarios (allowing borrowing pits or not) and clarifies the termination conditions of exchange (insufficient pits to meet the 3-pit rule), ensuring the completeness of logical derivation. 2. Calculationally accurate parts: The cumulative calculation of the number of peaches eaten is precise, such as 10 initial peaches + 3 exchanged peaches + 1 second exchanged peach = 14 peaches in the conventional scenario, and 14 + 1 borrowed-exchanged peach = 15 peaches in the optimized scenario. The calculation of leftover pits after each exchange is also error-free, like $10 - 9 + 3 = 4$, $4 - 3 + 1 = 2$, and $2 + 1 - 3 = 0$ (after borrowing), maintaining mathematical consistency. 3. Problem type adaptation strengths: The reasoning fully adapts to the recursive nature of the pit-peach exchange problem, repeatedly applying the core rule of "3 pits $\rightarrow$ 1 peach" and dynamically updating the number of pits and peaches. It also flexibly covers different boundary conditions of the problem (allowing borrowing or not), making the reasoning highly compatible with the problem's characteristics.", "answer": "15"



"reasoning": "1. Logically rigorous parts: Strictly tracks iterative pit exchanges, distinguishing used/leftover/new pits, clarifying exchange termination conditions, and covering both non-borrowing and borrowing scenarios without omissions. 2. Calculationally accurate parts: Precisely accumulates eaten peaches (14 in conventional scenario, 15 with borrowing) and correctly computes leftover pits after each exchange, maintaining mathematical consistency. 3. Problem type adaptation strengths: Repeatedly applies the "3 pits $\rightarrow$ 1 peach" rule to fit the problem's recursive nature and flexibly adapts to different boundary conditions."



"reasoning": "Start with 10 cents to buy 10 peaches. Each peach yields 1 pit, so 10 pits are obtained after eating all 10 peaches. Following the rule of 3 pits for 1 peach, exchange 9 pits for 3 peaches (1 pit leftover). Total peaches eaten: 10 + 3 = 13; total pits now: 1 leftover + 3 new pits = 4. Exchange another 3 pits for 1 peach (1 pit leftover again). Total peaches eaten: 13 + 1 = 14; total pits now: 1 leftover + 1 new pit = 2. To reach the maximum, borrow 1 pit to make 3 pits, exchange for 1 peach, and return the new pit after eating it. Total peaches eaten increases by 1.", "answer": "15"