Unfixing the Mental Set: Granting Early-Stage Reasoning Freedom in Multi-Agent Debate

Jing Wu, Suiyao Chen, Inseok Heo, Alexander Gutfraind,
Shengjie Liu, Chen Li, Bharathi Srinivasan, Xian Zhang, Michael Sharps

Abstract

Large language models (LLMs) have demonstrated remarkable performance across a wide range of tasks in recent years. While prior work has explored leveraging LLMs to generate synthetic data for self-improvement, repeated iterations often suffer from diminishing returns due to the reliance on homogeneous reasoning patterns and limited exploration of alternative perspectives. In this paper, we introduce a novel framework that enriches the reasoning process by encouraging critical thinking among multiple agents. Rather than deploying an ensemble of models with identical prompts, we propose a *strategy generator* that produces customized instructions tailored to each individual LLM. Acting as a critical thinking agent, the generator is iteratively fine-tuned using carefully selected strategies that are both diverse and effective. This approach fosters specialization within each model while promoting diversity across reasoning paths, enabling the system to maintain varied solution trajectories and achieve sustained performance gains through iterative refinement. We demonstrate the effectiveness of our method across a variety of agentic frameworks and complex reasoning tasks.

1. Introduction

In recent years, Large Language Models (LLMs) have experienced unprecedented advancements in domains such as language generation, comprehension, question answering, and translation (Touvron et al., 2023; Chowdhery et al., 2023; Achiam et al., 2023; OpenAI, 2024). These advancements are largely due to research efforts focused on the reasoning process (Wei et al., 2022; Wang et al., 2022; Yao et al., 2023; Besta et al., 2024; Gao et al., 2024). While logical chains have been significantly enhanced, LLMs continue to produce incorrect statements that conflict with their original claims. To address this issue, research on self-reflection has been proposed to improve consistency by evaluating and refining initial responses (Madaan et al., 2023; Kim et al., 2023; Shinn et al., 2023). However, the improvements become marginal with deeper self-reflection and multiple rounds of fine-tuning (Subramaniam et al., 2025). Additionally, constrained by homogeneous reasoning, these methods struggle to effectively correct mistakes. Without an external strategy as guidance, self-reflection ultimately leads to diminished performance (Huang et al., 2023).

One effective approach to addressing homogeneous reasoning is to encourage fine-tuning across multiple models using subsets of the dataset, promoting both specialization and diversification in responses. To ensure the quality of data for fine-tuning, a multi-agent debate (MAD) mechanism is employed to generate robust pseudo-labels (Du et al., 2023). An alternative way to conceptualize this issue is through the lens of a "mental set"—a cognitive bias that hinders the ability to explore diverse approaches, particularly when faced with novel or more complex tasks (Öllinger et al., 2008). Building on this observation, researchers have proposed the Diverse Multi-Agent Debate framework, which guides LLMs using predefined and varied reasoning strategies (Liu et al., 2015). As a result, the use of unique prompting strategies fosters divergent thinking and improves problem-solving capabilities.

Despite the success of prior approaches, we argue that predefined strategies are not always accessible and may fail to cover optimal solution paths. Moreover, customizing fine-tuning for each individual LLM agent incurs substantial computational overhead. To address these challenges, we propose to foster critical thinking through a novel *strategy generator* within a multi-agent debate framework—**Critical Thinking with Multi-Agent Debate (CMAD)**. The core innovation of the proposed method lies in treating problem-solving strategies as entirely undefined and fully optimizable. This unconstrained formulation enables broad exploration of the solution space, allowing the model to discover novel and potentially more effective strategies. However, the absence of structure introduces a high risk of failure

. Correspondence to: Jing Wu <jingwua@amazon.com >.

International Conference on Machine Learning Workshop (MAS), Vancouver, Canada. 2025.

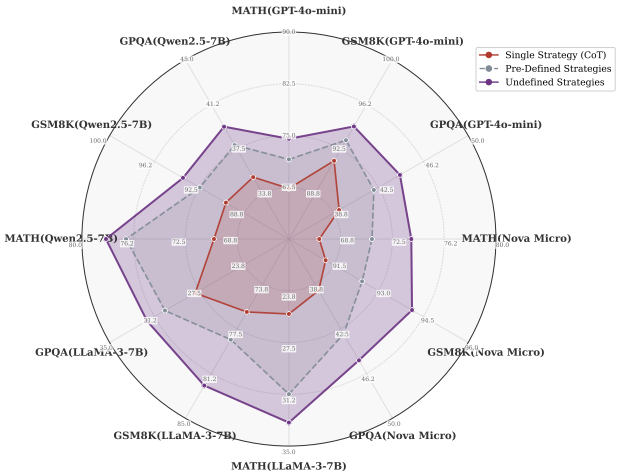

Figure 1. Undefined and optimizable strategies enhance problem-solving performance compared to both fixed single-strategy (Wei et al., 2022) and predefined-strategy baselines. Predefined strategies are instantiated from classic reasoning approaches including Chain-of-Thought Prompting, Step-Back Prompting, and Program-of-Thoughts Prompting, following prior paradigms.(Liu et al., 2015)

due to the stochastic nature of exploration. To mitigate this, we optimize the strategy space to converge toward a sweet point between exploration—encouraging diversity and creativity—and exploitation—ensuring solution quality and reliability.

More specifically, given a question, the strategy generator is encouraged to produce diverse yet undefined strategies for solving the problem. Each distinct strategy is assigned to an independent agent, which then generates a solution conditioned on its assigned strategy. These agents subsequently engage in critique and evaluation of each other's solutions following established debate frameworks (Du et al., 2023; Liu et al., 2015). While the initial strategies produced by the generator may be sub-optimal, we aim to iteratively improve its capability through a feedback loop grounded in two key perspectives: the `correctness` of final answers and the `diversity` of reasoning pathways. To assess correctness, we construct pseudo-labels by aggregating consensus outcomes from the multi-agent debate and multi-strategy evaluations. To measure diversity, we quantify the uniformity of the generated strategies. These two metrics are then used to guide sample selection, which in turn is used to fine-tune the strategy generator. This feedback mechanism fosters the emergence of novel, specialized reasoning strategies and drives continuous improvement in LLM performance.

We quantitatively validate the effectiveness of the approach across a diverse set of reasoning tasks and LLMs, demonstrating consistent performance gains. The framework is model-agnostic and integrates seamlessly with both open-source LLMs—such as `Qwen2.5` and `LLaMA-3`—and proprietary systems like `GPT-4o-mini` and `Nova Micro`, yielding marked improvements in solution quality. As shown in Figure 1, leveraging undefined and dynamically optimizable strategies within the critical thinking framework leads to significantly enhanced problem-solving capabilities, outperforming both fixed single-strategy baselines and predefined strategy paradigms. Furthermore, performance improves steadily with additional rounds of fine-tuning, demonstrating the scalability and robustness of the proposed framework. Our main contributions are summarized as follows:

- We propose a novel framework that encourages critical thinking in LLM agents by enabling them to generate diverse and undefined reasoning strategies, guided by a strategy generator.
- We introduce a comprehensive feedback loop that evaluates both the correctness and diversity of agent responses, providing reliable, dynamic, and specialized guidance to LLMs with minimal computational overhead.
- We empirically demonstrate that our fine-tuning paradigm for the strategy generator effectively encourages critical thinking and generalizes robustly across a wide range of datasets and popular LLMs.

2. Related Work

Multi-Agent LLM Reasoning: Multi-agent LLM reasoning enhances performance by enabling interaction and collaboration among multiple language model agents (Liang et al., 2023; Wang et al., 2023a; Khan et al., 2024; Chan et al., 2023). To facilitate more effective communication and coordination, prior work has explored role assignment strategies to specialize agent behaviors (Liang et al., 2023; Wang et al., 2023b; Chan et al., 2023). Another line of research encourages agents to challenge each other through iterative rounds of debate, promoting deeper reasoning and error correction (Du et al., 2023; Khan et al., 2024). While most debate-based frameworks treat agents as equally important participants, recent efforts have investigated expert-guided collaboration via meta-programming, inter-agent consistency, latent embeddings, and pre-defined reasoning paths (Hong et al., 2023; Xiong et al., 2023; Pham et al., 2023). However, existing approaches often overlook the heterogeneity of reasoning tasks, resulting in suboptimal performance of LLM solvers. Rather than adhering to a pre-defined problem-solving paradigm, we advocate for a dynamic critical thinking process that adapts strategies based on task demands.

Fine-tuning for Self-improvement: Fine-tuning has been widely adopted to improve the performance of LLMs (Welleck et al., 2022; Hsieh et al., 2023; Huang et al., 2022; Subramaniam et al., 2025; Zhang et al., 2024b). Considerable fine-tuning methods aim to optimize models using prior data to encourage strategy generation through self-iterated learning (Pang et al., 2024; Anthony et al., 2017; Polu et al., 2022; Parthasarathy et al., 2024). In addition, reinforcement learning has emerged as a popular self-training technique, often demonstrating better generalization (Chen et al., 2024a;a;b). Notably, most of these approaches rely on ground-truth data. In contrast, we diverge from these paths by employing unsupervised multi-agent interaction to achieve more consistent performance gains, following recent research (Subramaniam et al., 2025). More importantly, we emphasize the role of diversity in the sample selection process—an aspect that is frequently overlooked in prior work.

Critical Thinking for LLMs: Critical thinking is a powerful capability for promoting error correction and uncovering inconsistencies. It has been employed to detect noncompliance in statements (Kamath et al., 2020; Brahman et al., 2024), identify knowledge conflicts and misinformation, and reveal inconsistencies in problem framing (Xie et al., 2023; Zhou et al., 2023; Xu et al., 2023; Chen & Shu, 2024). These studies have significantly advanced LLMs' ability to recognize limitations, resolve contradictions, resist bias, and manage uncertainty. More recently, critical thinking has been adopted to enhance the reasoning capabilities of LLMs. This includes incorporating diverse reasoning paths, leveraging self-correction mechanisms (Tyen et al., 2023; Huang et al., 2023), quantifying reasoning quality through step-by-step scoring (Golovneva et al., 2022), and evaluating performance on specialized reasoning benchmarks (Zeng et al., 2024). Our work deviates from prior efforts by applying critical thinking to explore and enrich solution batches without pre-defined constraints on the reasoning process. This design enhances LLMs' ability to tackle more complex problems and improves the robustness of their outputs.

3. Method

While sticking to a single approach or a narrow set of strategies may lead to sub-optimal or dead ends, switching to unexplored methods can sometimes resolve challenging tasks more effectively. In this section, we present the framework for enabling diversified and critical thinking *without relying on pre-defined strategies*. We begin with an overview of the proposed approach in Section 3.1, illustrating how strategy generation guides the multi-agent system to complete a task. Section 3.2 then details the pipeline for extracting high-quality strategies from multi-agent interactions for fine-tuning.

3.1. Overall Framework

We present the overall workflow in Figure 2. Given a task x sampled from questions set $\mathcal{P}_q$ expressed in natural language, a strategy generator $G(x)$ takes x as input and proposes a set of high-level strategies that could potentially solve the task. We define this strategy generation process as:

$$S_1, S_2, \ldots, S_M = G(x), \qquad x \sim \mathcal{P}_q \tag{1}$$

Based on this, we denote the generated strategy set as $\mathbb{S} = \{\mathcal{S}_i \mid i = 1, 2, \ldots, M\}$. Correspondingly, we initialize M LLM agents, denoted as $\mathbb{A} = \{\mathcal{A}_i \mid i = 1, 2, \ldots, M\}$. In the first round of debate, each agent $\mathcal{A}_i$ is assigned a strategy $\mathcal{S}_i$ and tasked with generating a reasoning trajectory and a final answer, denoted as $y_{1,i}$, where the first subscript the subscript corresponds to the agent index and second indicates the debate round. Formally, the generation process in the first round is defined as:

$$y_{i,1} = \mathcal{A}_i(x; \mathcal{S}_i), \qquad i = 1, 2, \ldots, M. \tag{2}$$

In subsequent rounds, the responses and reasoning traces from the first round,i.e., $y_{1,1}, y_{2,1}, \ldots, y_{M,1}$, are aggregated into a shared historical context h_1, following the paradigm established in (Du et al., 2023). This shared history is then made available to all agents $\mathcal{A}i$. Conditioned on this history, the agents generate their second-round responses. This process is repeated iteratively in the following rounds. We define the general formulation as:

$$y_{i,n} = \mathcal{A}_i(x; h_{n-1}), \qquad i = 1, 2, \ldots, M, \quad n = 2, 3, \ldots, N. \tag{3}$$

3.2. Finetuning Strategy Generator with Selected Samples

To empower the strategy generator with critical thinking, we first carefully choose samples that actively guide the generator to solve the given problem x in correct and diverse ways. Correspondingly, we need metrics to quantify the correctness and diversity without knowledge of ground truth.

For correctness evaluation, we select the majority vote from the final round of debate across M agents and N rounds as the pseudo label $\hat{y}$. We then construct the dataset $\mathcal{D}_c$, consisting of samples whose solutions are aligned with $\hat{y}$, formally defined as:

$$\mathcal{D}_c \leftarrow \{y_{m,N} \mid y_{m,N} = \hat{y}, m \in \{1, 2, \ldots, M\}\} \tag{4}$$

While these pseudo labels are reliable with solution sharing among agents, the debate process inevitably leads to convergence toward similar reasoning trajectories. As a result, the final-round responses $y_{i,N}$ tend to follow closely aligned solution paths, producing increasingly similar outputs. This

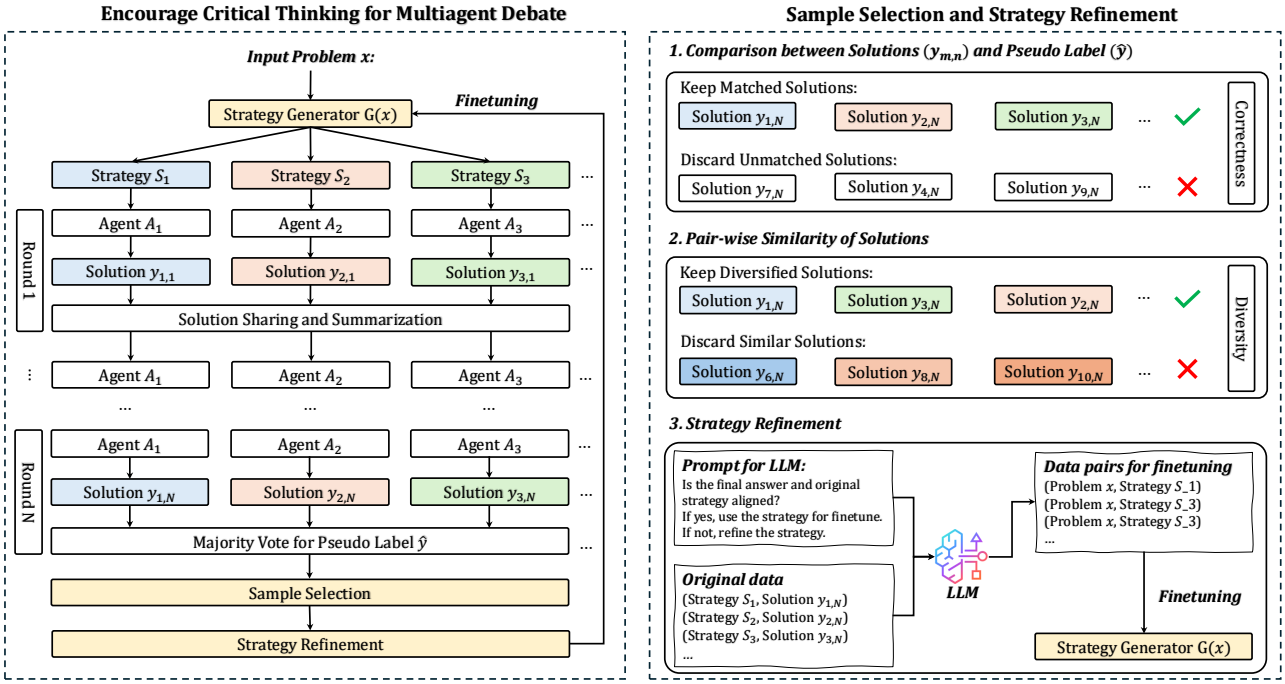

Figure 2. Overall Framework of Critical Thinking for Multi-Agent Debate. We first use strategy generation to guide each agent toward proposing diverse solutions. A majority vote is applied over the final-round answers to construct pseudo labels for fine-tuning (left). These pseudo labels help identify correct answers and effective strategies. Next, we evaluate the diversity of the generated solutions to further refine the pre-training data, improving the fine-tuning of the strategy generator and encouraging more diversified and reliable outputs (right). This figure illustrates a single fine-tuning iteration; applying multiple iterations can lead to further performance improvements.

convergence effect ultimately leads to diminishing returns in problem-solving performance as the number of agents and debate rounds increases.

To further encourage diverse critical thinking, we introduce a diversity metric that evaluates pairwise similarity among generated solutions. The metric is designed to satisfy three key properties: (1) asymptotic correctness—converging to a uniform distribution to ensure maximum diversity, (2) empirically effective with finite samples, and (3) the ability to capture non-linear semantic relationships. To this end, we adopt the Gaussian potential kernel (RBF kernel) (Cohn & Kumar, 2007; Borodachov et al., 2019), defined as:

$$G_t(u, v) = e^{-t\|u-v\|^2} = e^{-t(\|u\|^2+\|v\|^2-2u^\top v)}, \quad t > 0, \tag{5}$$

where $u, v \in \mathbb{R}^d$ are points in a d-dimensional Euclidean space, and t is a temperature parameter. Based on this, we define the diversity metric $\mathcal{D}$ as the logarithm of the expected pairwise Gaussian potential:

$$\begin{aligned}\mathcal{D}(f; t) &= \log \mathbb{E}_{x,y\sim\mathcal{P}_s}\left[G_t(f(x), f(y))\right] \\ &= \log \mathbb{E}_{x,y\sim\mathcal{P}_s}\left[e^{-t\|f(x)-f(y)\|^2}\right], \\ & t > 0,\end{aligned} \tag{6}$$

where x and y represent independent and identically distributed samples from $\mathcal{D}_c$, and f represents an embedding model.

Given the diversity metric, we proceed to select a subset of correct solutions. We start by including the first solution from our correct set $\mathcal{D}_c$, then systematically add solutions that are sufficiently different from those already selected. Specifically, for each candidate solution, we compute its pairwise similarity with all previously selected solutions and only include it if the maximum similarity is below a predefined threshold τ. This ensures each new addition contributes novel reasoning approaches. Formally, we construct our filtered dataset as:

$$\mathcal{D}_{div} = \{y_1\} \cup \{y_i \in \mathcal{D}_c \setminus \{y_1\} \mid \max_{y_j \in \mathcal{D}_{div}} |G_t(f(y_i), f(y_j))| > \tau\} \tag{7}$$

where τ is the similarity threshold that controls the diversity level of selected samples. A larger τ enforces greater diversity but potentially reduces the number of available training examples. This selection approach ensures that our strategy generator learns from a set of solutions that are not only correct but also represent diverse approaches to the same problem, thereby enhancing its critical thinking capabilities

across various reasoning paths.

Finally, we introduce an optional strategy refinement stage for alignment between the concrete solution from the last round of debate $y_{i,N}$ and initial strategy $\mathcal{S}_i$. Specifically, we introduce another strategy alignment agent $\mathcal{A}_{\text{align}}$, with prompt P_{ref} to identify and calibrate the logic of strategy $\mathcal{S}_i$ with the pseudo ground truth solutions $y_{i,N}$. The details of the prompt can be found in the Appendix. Formally, we define the process as:

$$\hat{S}_i = \mathcal{A}_{\text{align}}(S_i, P_{\text{ref}}; \hat{y}_i), \qquad i = 1, 2, \ldots, M. \tag{8}$$

Thereby, we are able to construct a high-quality dataset. For each question x, we apply $D_f(x, \hat{S}_i)$ for the fine-tuning task of the strategy generator. We illustrate the comprehensive summary of the procedures in Algorithm 1.

Algorithm 1 Critical Thinking Algorithm

Require: A set of input questions $\mathcal{P}_q = \{x_t\}$; The strategy generator $G(x)$; M model instances $\{\mathcal{A}_i | i = 1, 2, ..., M\}$; The number of debate rounds N; The number of finetuning iterations L; The diversity threshold τ

Initialize dataset $\mathcal{D}_f$ for finetuning
for $l = 1$ to L **do**
  for each x in $\mathcal{P}_q$ **do**
    for $j = 0$ to N **do**
      $S_1, \ldots, S_M \leftarrow G(x)$ {Generate Strategies (Eq. 1)}
      if $j = 0$ **then**
        $y_{1,1}, \ldots, y_{M,1} \leftarrow \mathcal{A}_1(x; \mathcal{S}_1), \ldots, \mathcal{A}_M(x; \mathcal{S}_M)$ {Generate Solutions (Eq. 2)}
      else
        $h_{j-1} \leftarrow$ Summarize the responses from agents in round $j-1$
        $y_{1,j}, \ldots, y_{M,j} \leftarrow \mathcal{A}_1(x; h_{j-1}), \ldots, \mathcal{A}_M(x; h_{j-1})$ {Refine Solutions (Eq. 3)}
      end if
    end for
    $\hat{y} \leftarrow$ Majority Vote of $\{y_{1,N}, \cdots, y_{M,N}\}$
    $\mathcal{D}_c \leftarrow$ Select samples aligned with $\hat{y}$ {Correctness Selection (Eq. 4)}
    $\mathcal{D}_{div} \leftarrow$ Select samples for diversity {Diversity Selection (Eq. 7)}
    $\hat{S}_1, \cdots, \hat{S}_M \leftarrow \mathcal{A}_{\text{align}}(S_1, P_{ref}), \cdots, \mathcal{A}_{\text{align}}(S_M, P_{ref})$ {Refine Strategies (Eq. 8)}
  end for
  $\mathcal{D}_f = \mathcal{D}_f \cup \{(x, \hat{S}_1), \cdots, (x, \hat{S}_M)\}$
  $G \leftarrow$ finetune $(G, \mathcal{D}_f)$
end for=0

4. Experiments

We evaluate the proposed method on three widely used benchmarks, as detailed in Section 4.1, and compare its performance against six recent strong baselines described in Section 4.2, as well as four representative large language models. The evaluated LLMs include two closed-source models—GPT-4o-mini by OpenAI (Achiam et al., 2023) and Amazon Nova Micro (Intelligence, 2024)—and two publicly available models—LLaMA-3-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024).

4.1. Benchmarks

MATH is a widely used benchmark comprising problems from high school mathematics competitions, spanning seven distinct subjects (Hendrycks et al., 2021). Accuracy is measured by comparing model predictions against ground truth answers, and correctness is determined through exact match.

GSM8K is a benchmark dataset comprising math word problems that require multi-step reasoning to solve (Cobbe et al., 2021). Each example consists of a problem statement, a corresponding numerical answer, and a step-by-step explanation. The problems primarily focus on basic arithmetic and introductory algebra.

GPQA is a challenging dataset consisting of 448 multiple-choice questions meticulously crafted by domain experts in biology, physics, and chemistry (Rein et al., 2024).

4.2. Baselines

We compare our proposed method against several state-of-the-art baselines. For our method, we employ three agents ($M = 3$), and for all debate-based approaches, we conduct two rounds of debate ($N = 2$). The baselines are as follows:

Chain-of-Thought Prompting (CoT) (Wei et al., 2022) This approach enables large language models to decompose complex problems by generating intermediate reasoning steps that lead to the final answer, thereby enhancing problem-solving capabilities through explicit step-by-step reasoning.

Step-Back Prompting (Zheng et al., 2023) This method improves reasoning by first prompting the model to abstract the problem to higher-level concepts and principles, then applying these abstractions to solve the original problem, effectively separating conceptual understanding from solution execution.

Multi-Agent Debate (Du et al., 2023) This framework facilitates multi-agent interaction where agents iteratively critique and refine solutions through structured debates, leveraging diverse perspectives to converge toward more robust

Method	GPT-4o-mini			Nova Micro		
	MATH	GSM8K	GPQA	MATH	GSM8K	GPQA
CoT	67.32	91.55	39.20	67.20	91.23	39.31
Step-Back Prompting	65.60	90.31	32.80	66.58	90.00	32.44
Multi-Agent Debate	70.57	92.91	40.36	70.34	92.44	41.23
Self-Reflection	67.75	90.25	39.28	66.74	91.25	38.62
Self-Contrast	62.43	90.13	37.93	63.76	90.18	36.57
DMAD	71.54	93.27	42.11	71.02	92.45	42.94
CMAD (Ours)	**74.52**	**94.42**	**44.29**	**73.87**	**94.12**	**45.15**
Method	**LLaMA-3-8B**			**Qwen2.5-7B**		
	MATH	GSM8K	GPQA	MATH	GSM8K	GPQA
CoT	25.43	76.10	27.84	70.43	90.27	35.18
Step-Back Prompting	24.87	75.31	24.43	69.52	88.29	33.52
Multi-Agent Debate	30.82	78.56	28.96	75.52	92.03	37.15
Self-Reflection	26.32	77.48	26.92	69.85	89.31	34.72
Self-Contrast	27.31	76.17	23.65	68.52	88.94	34.61
DMAD	31.24	78.42	30.37	76.80	92.46	37.86
CMAD (Ours)	**33.30**	**82.26**	**31.87**	**78.26**	**93.86**	**39.40**

Table 1. Quantitative comparison of the proposed method against six baseline approaches and four mainstream large language models. Best-performing scores are highlighted in gray.

answers.

Self-Reflection (Madaan et al., 2023) This technique enables models to evaluate and refine their initial outputs by critically analyzing their own reasoning, identifying potential errors or limitations, and generating improved solutions based on this introspection.

Self-Contrast (Zhang et al., 2024a)The proposed method generates diverse reasoning paths, identifies their discrepancies, and distills these differences into a structured checklist. The model then reflects on this checklist to iteratively revise each reasoning path, aiming to reach a coherent consensus.

DMAD (Liu et al., 2015) This method applies a set of predefined reasoning strategies to generate diverse solution paths, encouraging exploration of multiple problem-solving approaches to enhance solution quality and robustness.

4.3. Quantitative Results

We report the performance of the proposed method compared to six baseline approaches and four mainstream large language models. Notably, the strategy generator is fine-tuned for only a single iteration, i.e., $L = 1$. Results with additional refinement rounds are presented in Section 5.3.

As shown in Table 4.1, the proposed method consistently outperforms all baselines. The average improvement over the second-best method ranges from 1.2% to 9.8%. These results demonstrate that the proposed approach effectively generates feasible strategies to guide LLMs in solving complex questions.

5. Discussions and Visualizations

In this section, we present additional analyses and visualizations of results on `GPT-4o-mini` and `LLaMA-3-8B`. Specifically, we investigate the following questions: (1) What is the contribution of each component within the proposed framework? (2) How does the method perform with additional rounds of fine-tuning? (3) How does the diversity of solutions evolve across different stages of fine-tuning? (4) How does the threshold for diversity affect the overall performance of the framework?

5.1. Ablation Studies

We are interested in the variants of CMAD in the following settings:

CMAD with Pre-defined Strategy: We adopt three pre-

Method	GPT-4o-mini			LLaMA-3-8B		
	MATH	GSM8K	GPQA	MATH	GSM8K	GPQA
CMAD with Pre-defined Strategies	70.62	92.89	42.07	32.11	78.20	29.23
CMAD w/o Sample Selection	70.23	91.39	40.34	30.62	77.62	28.63
CMAD w/o Diversity Selection	71.03	91.54	41.74	30.77	78.04	30.04
CMAD w/o Correctness Selection	71.83	92.82	41.21	31.16	78.28	29.35
CMAD w/o Strategy Refinement	74.22	93.85	43.27	33.02	81.88	31.43
CMAD (Ours)	**74.52**	**94.42**	**44.29**	**33.30**	**82.26**	**31.87**

Table 2. **Ablation Results**: We analyze the contributions of individual components of the proposed method to overall performance.

defined reasoning strategies suggested by prior work (Liu et al., 2015) to generate fixed strategies for the guidance of solution generation.

CMAD w/o Sample Selection: All generated strategies are used for fine-tuning, regardless of their correctness or diversity.

CMAD w/o Correctness Selection: Only diverse strategies are used for fine-tuning, without explicitly matching them to pseudo labels.

CMAD w/o Diversity Selection: Strategies aligned with pseudo labels are used for fine-tuning, without explicitly enforcing diversity constraints.

CMAD w/o Strategy Refinement: Strategies aligned with pseudo-label diversity are used directly for fine-tuning, without any iterative refinement or revision.

The results are summarized in Table 4.3. We observe that fixed strategies guided by pre-defined templates yield similar results to previous research (Liu et al., 2015). Our experiments demonstrate that sample selection significantly improves the overall performance of the framework, suggesting that high-quality examples are critical for effective fine-tuning. Specifically, correctness-based selection produces gains by aligning the strategy generator with correct solutions. Although the improvement is modest, incorporating diversity further enhances performance. Moreover, strategy refinement contributes additional improvement by calibrating the initial strategies based on insights derived from alternative solutions.

5.2. Performance with Multiple Iterations of Finetuning

To verify the effectiveness of multiple iterations of fine-tuning, we report the performance of CMAD over five iterations in Figure 3. CMAD consistently improves the final results as the training iterations increase, with `GPT-4o-mini` varying from 74.52% to 76.80% and `LLaMA-3` from 33.30% to 37.04%. This improvement stems from the diverse samples selected purposefully.

In contrast, fine-tuning examples without considering diversity saturates after one iteration of fine-tuning and even begins to produce worse results. This observation is attributed to overfitting similar solutions, ultimately leading to the collapse of the training process. Additionally, we visualize the results with pre-defined strategy without fine-tuning as one of the baselines for comparison. We observe that overfitting with correct but less diverse samples could lead to worse results than pre-defined strategies, thereby further proving the importance of data quality and selection process.

5.3. Diversity with Multiple Iterations of Finetuning

While we encourage the strategy generator G to explore diverse strategies through simple prompt-level instructions, a single agent tends to converge toward generating similar—albeit correct—solutions after multiple rounds of fine-tuning. To better understand this behavior, we visualize the diversity of generated strategies, quantified by the uniformity metric defined in Equation 7. For more intuitive interpretation, we present the magnitude of uniformity in Figure 4. We observe that `CMAD` maintains diversity within a relatively stable range, in contrast to variants without diversity-based sample selection. For comparison, we also visualize the diversity of human-defined strategies, which serve as an idealized baseline for diversity. Since no fine-tuning is applied in these predefined settings, their diversity remains constant throughout.

5.4. Threshold of Diversity

Based on prior experiments, insufficient diversity leads to overfitting on semantically similar yet correct samples, as shown in Figure 3. However, setting the diversity threshold τ too high reduces the pool of eligible training samples, thereby constraining the capacity of the fine-tuned strategy generator. To better understand this trade-off, we analyze

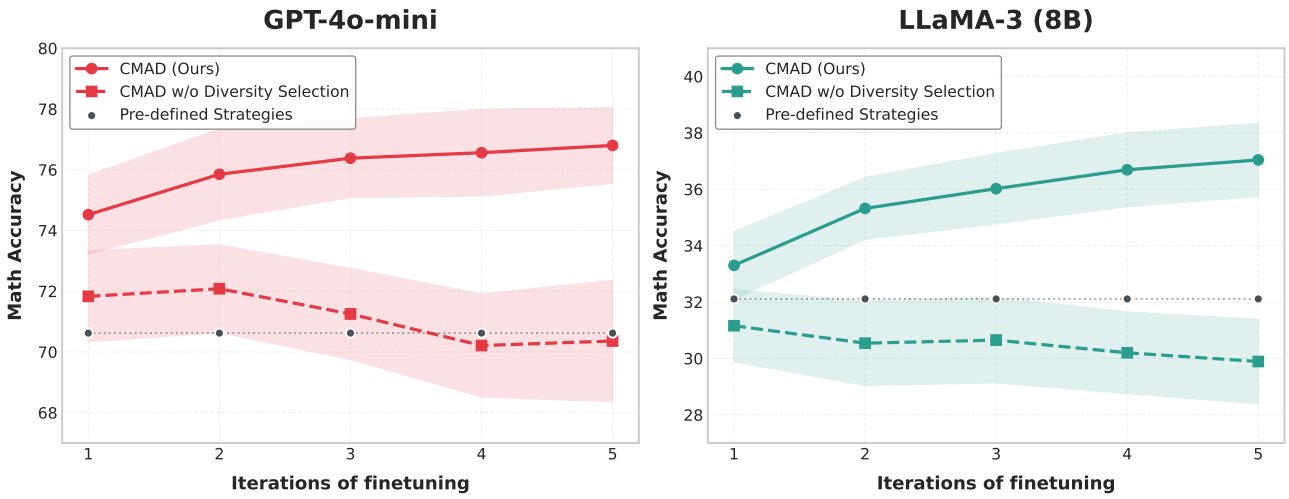

Figure 3. Critical thinking enhances the performance of LLMs on MATH datasets. While fine-tuning with carefully selected samples can improve reasoning capabilities, insufficient diversity in training examples ultimately degrades training effectiveness (dashed red lines), resulting in performance worse than debate methods with predefined strategies (dashed gray lines).

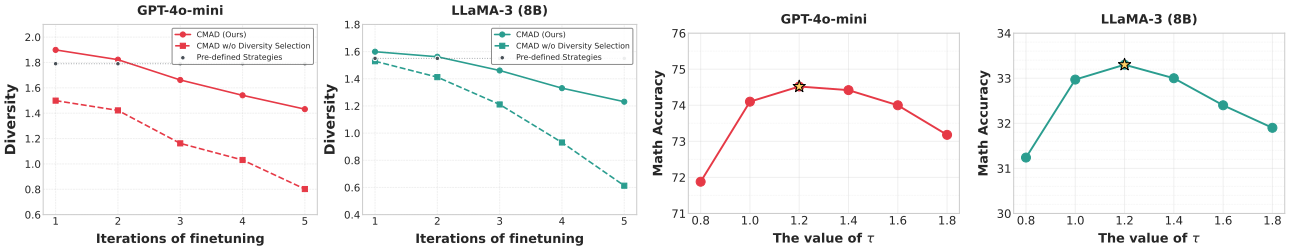

Figure 4. **Left:** Diversity under the critical thinking framework on the MATH dataset, measured via the uniformity magnitude (Equation 7). Diversity decline is mitigated with diversity-based sampling (solid red), while human-defined strategies remain stable due to the absence of fine-tuning (dashed gray). **Right:** CMAD accuracy across diversity thresholds τ using MATH, with optimal results near $\tau = 1.2$.

the effect of varying τ, as illustrated in Figure 4. When τ is low, model performance aligns with the baseline lacking diversity-based selection. As τ increases, CMAD performance improves, peaking at an optimal threshold before degrading due to insufficient training data. Notably, the effective range of τ increases with more capable LLMs, which naturally generate higher-quality and more diverse strategies.

6. Limitations and Conclusion

Limitations. Compared to prior work that primarily relies on prompt engineering, the proposed method inevitably incurs additional computational overhead due to the requirement for fine-tuning. Depending on the training approach—e.g., full supervised fine-tuning (SFT) or parameter-efficient methods like LoRA—the GPU memory requirement ranges from 8GB to 120GB. Additionally, achieving optimal results may require manual tuning of the diversity selection threshold, as the generative capacity and inherent diversity of strategies vary across different base LLMs.

Conclusion. In this paper, we introduced Critical Thinking with Multi-Agent Debate (CMAD), a novel framework that stimulates the latent creativity of LLMs by encouraging the generation of diverse and undefined solutions. By employing a strategy generator, the proposed method automatically equips multiple agents with distinct roles and reasoning pathways to collaboratively address and solve complex problems. We further introduce a feedback mechanism grounded in `Correctness` and `Diversity` to ensure the selection of high-quality solutions. These solutions are then used to fine-tune the strategy generator, promoting both creativity and reliability. Notably, CMAD enables autonomous self-improvement through iterative fine-tuning, achieving substantial performance gains without incurring heavy computational costs. The framework generalizes well across both closed-source and publicly available LLMs. We hope this work provides new insights into multi-agent debate, fine-tuning for self-correction, and the broader development of LLMs.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthony, T., Tian, Z., and Barber, D. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30, 2017.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Borodachov, S. V., Hardin, D. P., and Saff, E. B. *Discrete energy on rectifiable sets*, volume 4. Springer, 2019.

Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K., Hessel, J., et al. The art of saying no: Contextual noncompliance in language models. *Advances in Neural Information Processing Systems*, 37:49706–49748, 2024.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Chen, C. and Shu, K. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45 (3):354–368, 2024.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024a.

Chen, Z., Zhou, K., Zhao, W. X., Wan, J., Zhang, F., Zhang, D., and Wen, J.-R. Improving large language models via fine-grained reinforcement learning with minimum editing constraint. *arXiv preprint arXiv:2401.06081*, 2024b.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cohn, H. and Kumar, A. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Gao, P., Xie, A., Mao, S., Wu, W., Xia, Y., Mi, H., and Wei, F. Meta reasoning for large language models. *arXiv preprint arXiv:2406.11698*, 2024.

Golovneva, O., Chen, M., Poff, S., Corredor, M., Zettlemoyer, L., Fazel-Zarandi, M., and Celikyilmaz, A. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4): 6, 2023.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Intelligence, A. A. G. The amazon nova family of models: Technical report and model card. 2024.

Kamath, A., Jia, R., and Liang, P. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*, 2020.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R.,

Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

Kim, G., Baldi, P., and McAleer, S. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Liu, Y., Cao, J., Li, Z., He, R., and Tan, T. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*, 2015.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Öllinger, M., Jones, G., and Knoblich, G. Investigating the effect of mental set on insight problem solving. *Experimental psychology*, 55(4):269–282, 2008.

OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed: May 21, 2024.

Pang, R. Y., Yuan, W., He, H., Cho, K., Sukhbaatar, S., and Weston, J. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37: 116617–116637, 2024.

Parthasarathy, V. B., Zafar, A., Khan, A., and Shahid, A. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*, 2024.

Pham, C., Liu, B., Yang, Y., Chen, Z., Liu, T., Yuan, J., Plummer, B. A., Wang, Z., and Yang, H. Let models speak ciphers: Multiagent debate through embeddings. *arXiv preprint arXiv:2310.06272*, 2023.

Polu, S., Han, J. M., Zheng, K., Baksys, M., Babuschkin, I., and Sutskever, I. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344*, 2022.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Subramaniam, V., Du, Y., Tenenbaum, J. B., Torralba, A., Li, S., and Mordatch, I. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tyen, G., Mansoor, H., Cărbune, V., Chen, P., and Mak, T. Llms cannot find reasoning errors, but can correct them given the error location. *arXiv preprint arXiv:2311.08516*, 2023.

Wang, H., Du, X., Yu, W., Chen, Q., Zhu, K., Chu, Z., Yan, L., and Guan, Y. Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates. *CoRR*, 2023a.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023b.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Welleck, S., Lu, X., West, P., Brahman, F., Shen, T., Khashabi, D., and Choi, Y. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

Xie, J., Zhang, K., Chen, J., Lou, R., and Su, Y. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2023.

Xiong, K., Ding, X., Cao, Y., Liu, T., and Qin, B. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*, 2023.

Xu, R., Lin, B. S., Yang, S., Zhang, T., Shi, W., Zhang, T., Fang, Z., Xu, W., and Qiu, H. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*, 2023.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Zeng, Z., Liu, Y., Wan, Y., Li, J., Chen, P., Dai, J., Yao, Y., Xu, R., Qi, Z., Zhao, W., et al. Mr-ben: A comprehensive meta-reasoning benchmark for large language models. *arXiv e-prints*, pp. arXiv–2406, 2024.

Zhang, W., Shen, Y., Wu, L., Peng, Q., Wang, J., Zhuang, Y. T., and Lu, W. Self-contrast: Better reflection through inconsistent solving perspectives. In *Annual Meeting of the Association for Computational Linguistics*, 2024a. URL https://api.semanticscholar.org/CorpusID:266755862.

Zhang, Y., Khalifa, M., Logeswaran, L., Kim, J., Lee, M., Lee, H., and Wang, L. Small language models need strong verifiers to self-correct reasoning. *arXiv preprint arXiv:2404.17140*, 2024b.

Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., and Zhou, D. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*, 2023.

Zhou, W., Zhang, S., Poon, H., and Chen, M. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*, 2023.