

# CONTEXT LEARNING FOR MULTI-AGENT DISCUSSION

Xingyuan Hua<sup>1</sup>, Sheng Yue<sup>2,\*</sup>, Xinyi Li<sup>3</sup>, Yizhe Zhao<sup>1</sup>, Jinrui Zhang<sup>1</sup>, Ju Ren<sup>1,4</sup>

<sup>1</sup>Department of Computer Science and Technology, Tsinghua University,

<sup>2</sup>School of Cyber Science and Technology, Sun Yat-sen University,

<sup>3</sup>College of Computer Science, Northwest University,

<sup>4</sup>State Key Laboratory of Internet Architecture, Tsinghua University

## ABSTRACT

Multi-Agent Discussion (MAD) has garnered increasing attention very recently, where multiple LLM instances collaboratively solve problems via structured discussion. However, we find that current MAD methods easily suffer from discussion inconsistency—LLMs fail to reach a coherent solution—due to the misalignment between their individual contexts. In this paper, we introduce a multi-LLM context learning method (M2CL) that learns a context generator for each agent, capable of dynamically generating context instructions per discussion round via automatic information organization and refinement. Specifically, inspired by our theoretical insights on the context instruction, M2CL trains the generators to control context coherence and output discrepancies via a carefully crafted self-adaptive mechanism. It enables LLMs to avoid premature convergence on “majority noise” and progressively reach the correct consensus. We evaluate M2CL on challenging tasks, including academic reasoning, embodied tasks, and mobile control. The results show that the performance of M2CL significantly surpasses existing methods by 20%–50%, while enjoying favorable transferability and computational efficiency.<sup>1</sup>

## 1 INTRODUCTION

Large Language Models (LLMs) have demonstrated transformative impact across a large number of real-world domains, including education, healthcare, and scientific research, where an LLM instance is employed to automate the process of content generation, reasoning, or decision-making (Zheng et al., 2023; Imani et al., 2023; Zhang et al., 2024; Goyal et al., 2024; Wu et al., 2025; Yu et al., 2025). Yet, it has been recognized that single-LLM-built systems often struggle in the problems requiring complex multi-step reasoning or multi-tool using, such as complicated proof (Cobbe et al., 2021), large-scale code generation (Wang et al., 2025), and embodied agentic tasks (Ahn et al., 2022; Shen et al., 2024), because its single viewpoint of a problem easily limits the ability to explore multiple reasoning paths, leverage external tools effectively, or adapt to dynamic task requirements (Li et al., 2023; Smit et al., 2024).

Very recently, research has shifted towards *Multi-Agent Discussion* (MAD) (Smit et al., 2024), where multiple LLM instances collaboratively solve problems via structured discussions (Du et al., 2023; Liu et al., 2024). In a typical MAD framework, each LLM instance is pre-assigned a set of crafted *contexts* that represent diverse solution perspectives of the problem to be solved. Equipped with these distinct context instructions, LLMs continue to discuss with each other for a solution consensus (Park et al., 2023; Shanahan et al., 2023; Wei et al., 2023; Lu et al., 2024; Liu et al., 2024). Such ‘society-of-mind’ paradigms are expected to improve reasoning accuracy by enhancing creativity and expanding the search space for possible solutions, and have shown great potential across various complex tasks, including software engineering (Gu, 2023) and scientific discovery (Sprueill et al., 2024).

Albeit achieving improved performance over single-LLM settings, we find that current multi-agent collaboration approaches typically suffer from discussion inconsistency, that is, the majority of LLM instances fail to reach an agreement on a coherent solution (as showcased in Fig. 2), easily making the collaborative decision dominated by noise rather than principled reasoning. The underlying reason

\*Correspondence to: yuesh5@mail.sysu.edu.cn

<sup>1</sup>Code is available at <https://github.com/HansenHua/M2CL-ICLR26>.

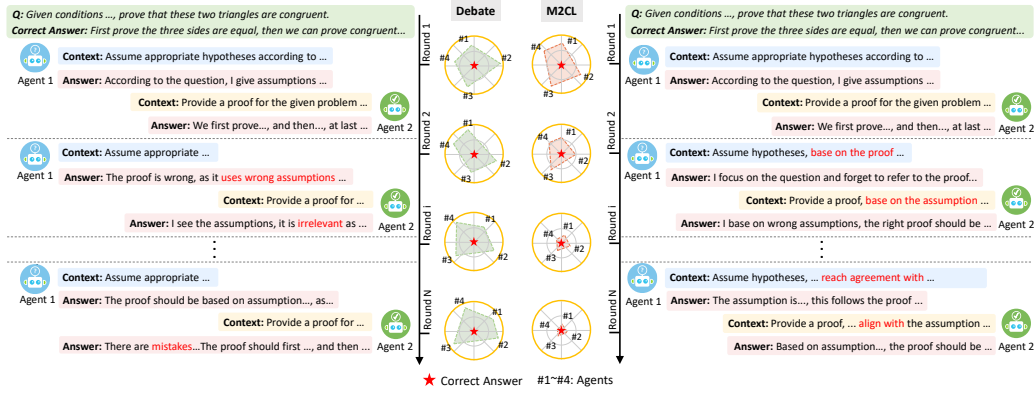


Figure 1: An illustration of context misalignment of an existing method (Debate Du et al. (2023)) on a multi-step proof task. Pre-assigned context instructions (in the blue and yellow boxes of the left part) provide insufficient guidance on information fusion, leading to conflict in reasoning.

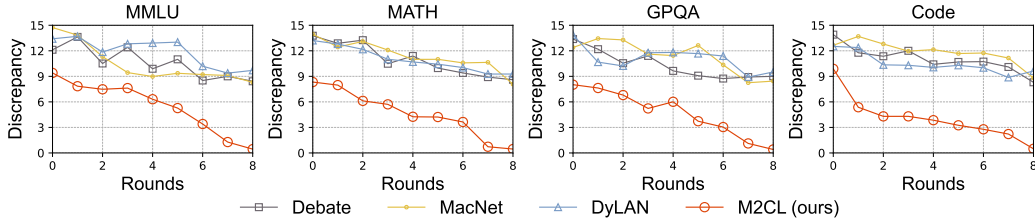


Figure 2: The discrepancy between the answers of participating LLM instances. The discrepancy is characterized by the maximum distance between participating LLMs’ output embeddings.

primarily lies in *context misalignment* between LLMs. On one hand, the pre-assigned contexts or role-based instructions lack a nuanced understanding of the task; they are often rigid, incomplete or biased, which would misguide the reasoning of individual LLMs (Jang et al., 2025). On the other hand, these contexts often fall short in effective fusion of information exchanged among LLMs, and thus hardly steer the discussion towards coherent solutions. As illustrated in Fig. 1, for a multi-step mathematical proof task, one LLM agent may correctly derive an intermediate result, yet another LLM—despite receiving this result in the extended context—does not effectively incorporate it into its own reasoning chain. Since the context instruction does not explicitly enforce leveraging conclusions from other LLMs, other LLMs may redundantly re-derive the step or even give an inconsistent argument (see Appendix H) for full results).

This paper aims to answer: “*how can we obtain the contexts for MAD that can continually guide multi-LLM discussion towards a correct consensus?*” A straightforward solution is to manually adapt instructions in contexts as the discussion progresses. It, however, is labor-intensive and requires extra expert knowledge, rendering it impractical for complex tasks or large-scale collaboration. Instead, a more reasonable approach is to develop context learning mechanisms that enable evolving the context instructions based on the intermediate discussion results. Although promising, it is highly challenging to evaluate the contribution of LLMs’ contexts to the final solution and control the coherence among inter- and intra-LLM outputs.

To tackle these challenges, we propose a *multi-LLM context learning* (M2CL) method for efficient MAD, which learns a context generator for each agent, capable of dynamically generating context instructions per discussion round via automatic information organization and refinement. First, we characterize the impact of initial and evolving contexts on the discussion performance. Building upon the analytical insights, we train the generators to control context coherence and output discrepancies. To strike the right tradeoff therein, we devise a self-adaptive balancing mechanism, enabling LLMs to progressively align on correct consensus while avoiding premature convergence on “majority noise”. Further, we develop a lightweight context initialization approach, which tends to assign LLMs with diverse initial instructions that are approximately orthogonal in the latent space, enabling sufficient coverage of complementary solution perspectives.

We systematically evaluate the proposed method across 9 challenging benchmarks, including LLM reasoning, embodied agentic tasks, and mobile GUI control. The results demonstrate that `M2CL` significantly enhances the MAD performance under various numbers of participating LLMs, consistently outperforming existing methods by 20%–50%, particularly in complex GUI control tasks. `M2CL` enjoys a more favorable MAD “scaling law” and exhibits great efficiency, where a runtime overhead of at most 10% suffices to achieve more than 20% performance gains. Further, we find that the learned context generators can be migrated to different LLM architectures with consistent performance improvement.

Our main contributions are summarized as follows:

- We propose `M2CL`, a principled multi-LLM context learning method that learns a context generator for each agent, capable of dynamically generating context instructions per discussion round.
- We devise a lightweight context initialization approach that can assign LLMs with diverse initial instructions, thereby enabling sufficient coverage of complementary solution perspectives.
- We systematically evaluate the proposed method across a range of challenging benchmarks and corroborate the efficacy and efficiency of the proposed method.

## 2 RELATED WORK

**Multi-agent framework.** Multi-agent frameworks have been proposed to enhance the reasoning capabilities of single-LLM prompting methods. Several works utilize two LLMs to iteratively generate and evaluate to refine the final answer (Raman et al., 2022; Yao et al., 2023; Paul et al., 2023; Madaan et al., 2024). However, scaling them to a larger number of LLMs to make full use of the wisdom of crowd is challenging due to the pairwise dependency of generation and evaluation. To incorporate more LLMs, several studies have explored another multi-agent framework in which each LLM has access to the history of all preceding LLMs’ responses (Chan et al., 2023; Du et al., 2023; Liang et al., 2023; Smit et al., 2024; Liu et al., 2024; Zhuge et al., 2024). Most existing methods remain constrained by manually defined inter-LLM topologies or workflows. More recently, researches have attempted to overcome this limitation by automatically optimizing workflows (Zhuge et al., 2024; Liu et al., 2024; Zhang et al., 2025b). However, such evolution is often one-shot. For instance, Zhuge et al. (2024) describe language agent systems as optimizable computational graphs and enables automatic improvements of LLM prompts and inter-LLM orchestration. Zhang et al. (2025b) employs Monte Carlo Tree Search to construct complex Multi-LLM systems tailored to a specific task domain. Liu et al. (2024) propose a multi-LLM collaboration method which constructs dynamic communication structure by scoring other LLMs. While these approaches allow for scaling the number of LLMs, designing appropriate contexts for them to collaboratively solve problems remains difficult, as single-viewpoint context instructions fall short in guidance on how to organize and refine information across agents.

**Context learning.** Context learning has recently attracted significant attention, as it allows LLMs to adjust their behavior at inference time by modifying the input context, without requiring any gradient-based training (Von Oswald et al. (2023); Todd et al. (2024); Li et al. (2024a)). One line of work focuses on context selection, which identifies the most relevant examples or information to include in the context (Zhang et al. (2022); Lu et al. (2023); Xiong et al. (2024); Purohit et al. (2025)). To eliminate manual selection, Lu et al. (2023) leverage reinforcement learning to learn a context selection policy. Xiong et al. (2024) query LLM to obtain knowledge and then query a retriever to obtain the final context. Beyond selection, some recent methods have proposed to generate or evolve context (Zhuge et al., 2024; Li et al., 2024b; Zhang et al., 2025a). Madaan et al. (2024) leverage feedback generated by LLMs as extra prompt and iteratively incorporate it into revised drafts, aiming to enhance the coherence of the generated text. Pandita et al. (2025) dynamically refines prompts during inference using textual feedback from prior outputs, aiming to improve contextual alignment and generate more consistent responses. Albeit with promising results, these methods are grounded in single-LLM formulations and struggle with inter-LLM inconsistency during multi-agent discussion, where it is crucial to guide LLMs to make full use of others’ intermediate results.

## 3 PRELIMINARIES

In this section, we provide necessary backgrounds and definitions of our investigated problem.

**Context learning.** To formally define context learning, we begin with the standard probabilistic model of an autoregressive LLM. An autoregressive model, parameterized by  $\phi$ , generates an output sequence  $X = (x_1, x_2, \dots, x_L)$  given an input context  $C$  by maximizing the conditional probability:

$$P_\phi(X|C) = \prod_{l=1}^L P_\phi(x_l|x_{<l}, C). \quad (1)$$

Historically, in the paradigm of prompt engineering, the context  $C$  was treated as a static string of text. This view is insufficient for complex agentic tasks which leverage dynamic, structured, and multifaceted information stream. To address this, context learning is redefined as a dynamically organized collection of information, denoted as  $\{c_1, c_2, \dots, c_n\}$ . Each component  $c_k$  can be instructions, external knowledge (Lewis et al., 2020), available external tools (Qin et al., 2023), and memory (Zhang et al., 2025a). These components are sourced, filtered, and formatted by a set of functions  $f_k$ , and orchestrated into a coherent representation by a high-level assembly function  $\mathcal{A}$ :

$$C = \mathcal{A}(f_1(c_1), f_2(c_2), \dots, f_n(c_n)). \quad (2)$$

**Multi-LLM context learning.** MAD involves a set of  $N$  LLM instances collaboratively solving a task via multi-round inter-LLM discussion (debate). Each LLM  $i$  is endowed with an evolving instruction context in terms of the task description, available external tools and current knowledge aggregation. At the  $t$ -th round, we use the concatenation function as the assembly function  $\mathcal{A}$  and construct the context with three components: (i) the task goal  $P \in \mathbb{R}^{d_{model} \times n}$ , (ii) the concatenation of responses from all other LLMs in the previous round  $\bar{X}_i^{t-1} = [X_j^{t-1}]_{j \neq i} \in \mathbb{R}^{d_{model} \times n}$ , where  $X_j^{t-1}$  is the response of LLM  $j$  at round  $t-1$ , and (iii) the current instruction context  $I_i^t \in \mathbb{R}^{d_{model} \times n}$  ( $d_{model}$  refers to the dimension of embedding). The task goal represents the ultimate objective of collaboration and remains invariant across rounds. The second component serves as a dynamic memory that incorporates cross-LLM interaction history. For the third component, in contrast to existing efforts relying on static preassigned roles, we employ an instruction generator  $\mathcal{G}$ , parameterized by  $\theta_i$ , to adaptively refine it into a per-step instruction  $I_i^t$ , conditioned on the task goal  $P$ , the initial instruction  $I_i^b$ , and the concatenated response  $\bar{X}_i^{t-1}$ :

$$I_i^t = \mathcal{G}_{\theta_i}([P; I_i^b; \bar{X}_i^{t-1}]). \quad (3)$$

Given context  $C_i^t = [I_i^t; \bar{X}_i^{t-1}; P]$ , each LLM  $i$ , parameterized by  $\phi_i$ , generate its response by:

$$X_i^t = \arg \max_X P_{\phi_i}(X|C_i^t). \quad (4)$$

After  $T$  rounds of interaction, the final result is obtained by a majority vote on the LLMs' outputs generated in the final round.

## 4 MOTIVATION

In this section, we investigate the quantified impact of the contexts on the MAD performance. First, we introduce the formulation of attention activation, denoted as  $a(\cdot)$ , as follows (Vaswani, 2017):

$$a(C_i^t) \doteq W_V[I_i^t; \bar{X}_i^{t-1}; P] \text{softmax} \left( \frac{(W_K[I_i^t; \bar{X}_i^{t-1}; P])^T W_Q P}{\sqrt{d}} \right) \quad (5)$$

where  $\sqrt{d}$  is a scaling factor.  $W_Q$ ,  $W_K$ , and  $W_V$  denote the parameter weight matrices in the attention mechanism.  $a(C_i^t) \in \mathbb{R}^{d_{model} \times n}$  is a matrix with the same shape as  $[P; X; I]$

From this, we have the following theorem characterizing the total distance between the activations of the correct answer  $a_c$  and that induced by context  $C_j^t$ . Here, we utilize activation distance instead of token embedding distance as attention activation captures deep representational similarity learned through the model's internal reasoning process, making it more robust to superficial linguistic variations.

**Theorem 4.1.** Assume that the attention activation function is  $L_a$ -smooth and define the weight vector as  $\omega \doteq [\omega_1, \omega_2, \dots, \omega_N]$ , with the initial context  $C_i^b \doteq [I_i^b; P]$ . Then, the following fact holds:

$$\begin{aligned} \sum_{i=1}^N \|a_c - a(C_i^t)\| &\leq \sum_{i=1}^N \left( \sum_{j=1}^N \|a(C_i^t) - a(C_j^t)\| + (N+1)L_a \|C_i^t - C_i^b\| \right) \\ &\quad + N \min_{\omega} \|a_c - \sum_{i=1}^N \omega_i a(C_i^b)\|. \end{aligned} \quad (6)$$

*Proof.* For a detailed proof, please refer to Appendix B.  $\square$

Theorem 4.1 corroborates the necessity of multi-LLM context learning, consisting of both initialization and evolution. The first term,  $\|a(C_i^t) - a(C_j^t)\| + (N+1)L_a \|C_i^t - C_i^b\|$ , captures the divergence among LLMs’ activations as well as the deviation from their initial contexts. This indicates contexts must be continuously evolved to reduce inter-LLM discrepancies while keeping coherent, promoting consistency of reasoning chains. The second term,  $\min_{\omega} \|a_c - \sum_{i=1}^N \omega_i a(C_i^b)\|$ , depends solely on the initial contexts  $I_i^b$ . It implies that orthogonality among initial activations provides a comprehensive basis, allowing the contexts to approximate the correct activation more effectively. This highlights the importance of diverse solution perspectives at the initialization stage. Together, the result motivates a two-stage design: initializing contexts to ensure diverse solution perspectives and evolving intermediate contexts to drive a consensus across LLMs.

## 5 MULTI-LLM CONTEXT LEARNING

In this section, we introduce a multi-LLM context learning method that can select appropriate contexts for each LLM and dynamically adapt the context according to the evolving task completion status.

### 5.1 CONTEXT INITIALIZATION

The context initialization serves as the foundation of the entire multi-agent interaction process, as it frames the capability scope of LLMs and fundamentally shapes the chain of subsequent information exchange. Therefore, we propose to select initial contexts from a predefined pool  $\{I_i^b\}_{i=1}^M$  containing prompts with diverse perspectives.

Motivated by Theorem 4.1, we provide the following context initialization mechanism:

$$\mathbf{I}^b = \{I_1^b, \dots, I_N^b\} = \arg \min_{\mathbf{I}^b} \left\{ \min_{\omega} \left\| \sum_{i=1}^N \omega_i a([I_i^b; P]) - a_c \right\| \right\}. \quad (7)$$

Eq. (7) identifies a subset of contexts  $\mathbf{I}^b$  whose activation best reconstruct the target activation  $a_c$ . As the dimension of the activation matrix  $a([I_i^b; P]) \in \mathbb{R}^{d_{model} \times n}$  is far larger than the number of selected contexts  $N$ , a set of matrices that aims to best reconstruct the correct activation  $a_c$  naturally tends toward forming a set of basis-like directions. The resulting near-orthogonal activations form a compact basis, ensuring each context contributes unique, non-overlapping information for subsequent discussion.

However, Eq. (7) is impractical since the correct activation  $a_c$  is not accessible during initialization. Inspired by prior work (Yang et al., 2024b; 2025), we project the activation into a latent space with function  $f$ :

$$\mathbf{I}^b = \{I_1^b, \dots, I_N^b\} = \arg \min_{\mathbf{I}^b} \left\{ \min_{\omega} \left\| \sum_{i=1}^N \omega_i f(a([I_i^b; P])) - f(a_c) \right\| \right\}, \quad (8)$$

Then, we use the problem space as the selected contexts are solely dependent on the problem. This projection preserves the original orthogonality properties of activations, ensuring that diverse perspectives remain distinguishable in other spaces. Therefore, we reformulate the initialization

mechanism:

$$\mathbf{I}^b = \{I_1^b, \dots, I_N^b\} = \arg \min_{\mathbf{I}^b} \left\{ \min_{\omega} \left\| \sum_{i=1}^N \omega_i f(a([I_i^b; P])) - v_P \right\| \right\}, \quad (9)$$

where  $v_P$  denotes the sentence vector of the question  $P$ . To obtain the projection  $f(\cdot)$  (parameterized by  $\phi_f$ ), we utilize the activation of inputting the answer  $A$  with the question  $P$  as the correct activation  $a_c = a([A; P])$  and provide the loss function as:

$$L(\phi_f) = \|v_P - f(a([A; P]))\|. \quad (10)$$

However, directly leveraging Eq. (9) to initialize context is computational costly as it requires all the context activations in the context pool. Therefore, we distill a lightweight  $\mathcal{F}(\cdot)$  (parameterized by  $\phi_F$ ) which directly projects the initial context into the problem space through the following loss function:

$$L(\phi_F) = \|\mathcal{F}([I_i^b; P]) - f(a([I_i^b; P]))\|. \quad (11)$$

Then, we provide the final context initialization as:

$$\mathbf{I}^b = \{I_1^b, \dots, I_N^b\} = \arg \min_{\mathbf{I}^b} \left\{ \min_{\omega} \left\| \sum_{i=1}^N \omega_i \mathcal{F}([I_i^b; P]) - v_P \right\| \right\}. \quad (12)$$

Eq. (12) provides a computational efficient context initialization (as illustrated in Fig. 23) that aligns with Eq. (7) to encourage orthogonality of selected context activations, thereby providing diverse reasoning perspectives and expanding the search space for solutions.

## 5.2 CONTEXT EVOLUTION

As discussed in Section 5.1, context initialization aims to select contexts with diverse perspectives to solve the problem. Nevertheless, these individual contexts address the problem from a single viewpoint while lacking explicit instructions on how to incorporate perspectives provided by other LLMs. Directly using them often results in inconsistent reasoning across LLMs due to misaligned inter-LLM guidance. To address this, we iteratively refine the instruction to integrate collaborative instructions that guide the LLMs more effectively toward generating the desired solution.

### 5.2.1 EVALUATE THE CONTRIBUTION OF CONTEXTS

To achieve efficient context evolution, it is crucial to quantify the contribution of different contexts (refer to a utility function). A natural solution is to assign utilities only at the final round. However, this leads to inefficient and unstable training process due to its sparsity Liu et al. (2025). On the other hand, current methods typically use correctness as the only criterion (Zelikman et al., 2022; Rafailov et al., 2023; Guo et al., 2025). Albeit with promising performance in single-LLM question answering, this simple criterion ignores the complex dependency among LLMs with different contexts in MAD. As a result, an LLM that provides crucial insights which enable others to reach the correct answer may still receive a low utility, simply because it fails to produce the correct final answer itself. To tackle these challenges, we introduce a novel round-wise criterion to evaluate the contribution of  $C_i^t$  in multi-agent interaction:

$$\max_{j \in [N]} \left\{ -\alpha \|C_i^t - C_j^b\| - \|a(C_i^t) - a(C_j^t)\| \right\}, \quad (13)$$

where  $\alpha \geq 0$  is a weighting parameter and  $\|a(C_i^t) - a(C_j^t)\|$  represents the activation difference between contexts.

**Remarks.** Of note, if removing the second term of Eq. (13), the context generation reduces to that of prompt engineering:  $\min_{\mathbf{I}^t} \sum_{i \in [N]} \alpha \|C_i^t - C_i^b\|$ , which utilize fixed instructions in contexts to LLMs throughout the discussion (Liu et al., 2024; Lu et al., 2024). As mentioned in Section 1, this easily results in inconsistency across LLMs. Thus, we introduce  $\|a(C_i^t) - a(C_j^t)\|$  into the objective to encourage LLMs to remain aligned with each other.

This design addresses the two aforementioned challenges. First, this round-wise criterion provides denser feedback, mitigating the inefficiency and instability caused by sparse final-round utilities.

Second, the activation-based alignment term captures inter-LLM dependencies, thereby encouraging them to offer guidance on how to integrate others' responses.

Applying Eq. (13) as the criterion for LLM  $i$  requires the contexts of all other LLMs. However, since these contexts are simultaneously being optimized, a bias inevitably arises as each LLM is updated based on stale snapshots of others' contexts. Such outdated information introduces a biased utility, which in turn hinders consistency and degrades the quality of the learned contexts. To overcome this issue, we decouple the inter-LLM dependencies by designing an alternative per-LLM criterion:

$$-\alpha\|C_i^t - C_i^b\| - \|a([I_i^t, P]) - a([X_i^{t-1}, P])\|. \quad (14)$$

In Lemma C.1, we can prove that the summation of Eq. (14) over all LLMs serves as an upper bound of the summation of Eq. (13) over all LLMs. Intuitively, this criterion conveys a clear meaning: the first term,  $-\alpha\|C_i^t - C_i^b\|$ , preserves the fundamental problem-solving capability endowed by the initialization, while the second term,  $-\|a([I_i^t, P]) - a([X_i^{t-1}, P])\|$ , enforces consistency between an LLM's current instruction and its own previous response. This local consistency constraint implicitly aligns all LLMs, because when every LLM evolves its context in a temporally coherent way, the divergence among them is gradually reduced. Hence, the collective effect of all LLMs following this rule is that their contexts evolve coherently across rounds, progressively shrinking activation differences and ultimately driving the MAD toward consistency in their final answers.

To intuitively show the tightness of the bound, consider the case where  $I_i^t = X_i^{t-1}$  holds, i.e., the instruction of each LLM in the next round equals its response in the previous round. That means  $a(I_i^t) = a(X_i^{t-1})$ . Given that all LLMs receive the same context, the second term of both Eq. (13) and Eq. (14) becomes zero. In this case, the two formulations coincide.

### 5.2.2 MULTI-ROUND CONTEXT EVOLVING

We slightly abuse notation by denoting  $a(I_i^t) = a([I_i^t; P])$  and  $a(X_i^{t-1}) = a([X_i^{t-1}; P])$ , and provide the multi-round context evolution objective by accumulating contributions as follows:

$$\sum_{t=1}^T \max_{I_i^t} \left\{ -\alpha\|C_i^t - C_i^b\| - \|a(I_i^t) - a(X_i^{t-1})\| \right\}. \quad (15)$$

However, selecting the optimal value for the weight  $\alpha$  is non-trivial, as it must be carefully tuned for the evolving discussion process. To alleviate the need for manual tuning, we recast Eq. (15) into a constrained optimization problem, where the context adjustment  $\|I_i^t - I_i^b\|$  is treated as a constraint:

$$\begin{aligned} & \min_{I_i^1} \left( \|a(I_i^1) - a(X_i^0)\| + \min_{I_i^2} \left( \|a(I_i^2) - a(X_i^1)\| + \dots + \min_{I_i^T} \|a(I_i^T) - a(X_i^{T-1})\| \right) \right) \\ & \text{s.t. } \|C_i^t - C_i^b\| \leq \beta, \quad \forall t, i. \end{aligned} \quad (16)$$

According to the detailed derivation in Appendix E, we prove that taking the dual of this problem recovers Eq. (15) and yields an auxiliary update for the dual variable  $\alpha$ . Consequently, the optimization of Problem (16) can be implemented using an approximate dual gradient descent procedure, alternating gradient updates of  $L(\theta_i)$  and  $L(\alpha_i)$ .

$$\begin{aligned} L(\theta_i) &= \|a(\mathcal{G}_{\theta_i}(P, I_i^b, \bar{X}_i^{t-1})) - a(X_i^{t-1})\| + \alpha\|C_i^t - C_i^b\| \\ L(\alpha_i) &= \alpha_i(\beta - \|\mathcal{G}_{\theta_i}(P, I_i^b, \bar{X}_i^{t-1}) - C_i^b\|). \end{aligned} \quad (17)$$

At the beginning of the discussion, the initial contexts are intentionally diverse to encourage multi-perspective reasoning of the problem. At this stage, when the answers differ greatly,  $\alpha$  decreases rapidly, thereby weakening the constraint on the distance between the generated contexts and their initial contexts. This guides the generation of contexts towards promoting faster convergence on a unified solution. As the discussion progresses and the LLMs gradually reach agreement,  $\alpha$  will be kept at a certain level. This adjustment prevents premature consensus and supports a richer, more comprehensive final solution by shifting the contexts' focus toward preserving multiple perspectives and exploring nuanced differences.

Overall, we name our proposed algorithm M2CL, with its pseudocode detailed in Appendix F.4.

Table 1: Accuracy (%) on different datasets. The number of LLMs is 4 for all dataset. We exhibit the performance advantage with BoN and highlight the **best** result.

Model	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
Qwen-7B	Single	61.2 <sub>±13.0</sub>	12.9 <sub>±12.0</sub>	20.2 <sub>±16.2</sub>	51.2 <sub>±11.3</sub>	23.8 <sub>±7.7</sub>	25.2 <sub>±10.1</sub>	15.6 <sub>±5.5</sub>	21.0 <sub>±5.3</sub>
	BoN	74.2 <sub>±0.0</sub>	24.9 <sub>±0.0</sub>	36.4 <sub>±0.0</sub>	62.5 <sub>±0.0</sub>	31.5 <sub>±0.0</sub>	35.3 <sub>±0.0</sub>	21.1 <sub>±0.0</sub>	26.3 <sub>±0.0</sub>
	Debate	71.1 <sub>±3.1</sub>	19.9 <sub>±5.0</sub>	28.7 <sub>±7.7</sub>	60.0 <sub>±2.5</sub>	30.6 <sub>±0.9</sub>	32.2 <sub>±3.1</sub>	21.0 <sub>±0.1</sub>	24.8 <sub>±1.5</sub>
	DyLAN	74.3 <sub>±0.1</sub>	26.7 <sub>±1.8</sub>	35.4 <sub>±1.0</sub>	63.4 <sub>±0.9</sub>	29.8 <sub>±1.7</sub>	29.4 <sub>±5.9</sub>	18.4 <sub>±2.7</sub>	23.4 <sub>±2.9</sub>
	GPTSwarm	76.3 <sub>±2.1</sub>	26.2 <sub>±1.3</sub>	35.6 <sub>±0.8</sub>	62.7 <sub>±0.2</sub>	29.9 <sub>±1.6</sub>	30.8 <sub>±4.5</sub>	20.6 <sub>±0.5</sub>	24.5 <sub>±1.8</sub>
	MacNet	71.5 <sub>±2.7</sub>	21.4 <sub>±3.5</sub>	30.8 <sub>±5.6</sub>	59.3 <sub>±3.2</sub>	33.6 <sub>±2.1</sub>	37.0 <sub>±1.7</sub>	21.5 <sub>±0.4</sub>	29.5 <sub>±3.2</sub>
	M2CL (ours)	92.5 <sub>±18.3</sub>	47.8 <sub>±22.9</sub>	66.1 <sub>±29.7</sub>	80.3 <sub>±17.8</sub>	39.9 <sub>±8.4</sub>	45.3 <sub>±10.0</sub>	33.6 <sub>±12.5</sub>	34.7 <sub>±8.4</sub>
Qwen-14B	Single	67.2 <sub>±12.5</sub>	21.6 <sub>±6.2</sub>	21.2 <sub>±11.6</sub>	56.7 <sub>±12.7</sub>	30.1 <sub>±7.4</sub>	31.3 <sub>±10.3</sub>	18.7 <sub>±6.9</sub>	25.7 <sub>±2.8</sub>
	BoN	79.7 <sub>±0.0</sub>	27.8 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	69.4 <sub>±0.0</sub>	37.5 <sub>±0.0</sub>	41.6 <sub>±0.0</sub>	25.6 <sub>±0.0</sub>	28.5 <sub>±0.0</sub>
	Debate	77.2 <sub>±2.5</sub>	27.4 <sub>±0.4</sub>	30.3 <sub>±2.5</sub>	66.9 <sub>±2.5</sub>	36.8 <sub>±0.7</sub>	39.4 <sub>±2.2</sub>	26.4 <sub>±0.8</sub>	30.0 <sub>±1.5</sub>
	DyLAN	86.8 <sub>±7.1</sub>	31.2 <sub>±3.4</sub>	39.4 <sub>±6.6</sub>	76.6 <sub>±7.2</sub>	34.8 <sub>±2.7</sub>	36.0 <sub>±5.6</sub>	22.8 <sub>±2.8</sub>	28.0 <sub>±0.5</sub>
	GPTSwarm	87.0 <sub>±7.3</sub>	31.3 <sub>±3.5</sub>	38.7 <sub>±5.9</sub>	75.9 <sub>±6.5</sub>	34.9 <sub>±2.6</sub>	35.9 <sub>±5.7</sub>	25.0 <sub>±0.6</sub>	28.4 <sub>±0.1</sub>
	MacNet	78.9 <sub>±0.8</sub>	28.7 <sub>±0.9</sub>	31.9 <sub>±0.9</sub>	70.4 <sub>±1.0</sub>	39.2 <sub>±1.7</sub>	46.0 <sub>±4.4</sub>	32.2 <sub>±6.6</sub>	32.7 <sub>±4.2</sub>
	M2CL (ours)	93.7 <sub>±14.0</sub>	51.7 <sub>±23.9</sub>	66.2 <sub>±33.4</sub>	91.1 <sub>±21.7</sub>	48.2 <sub>±10.7</sub>	56.1 <sub>±14.5</sub>	42.0 <sub>±16.4</sub>	43.0 <sub>±14.5</sub>
Qwen-72B	Single	72.5 <sub>±11.7</sub>	31.6 <sub>±19.4</sub>	34.9 <sub>±11.0</sub>	59.1 <sub>±13.1</sub>	48.2 <sub>±9.3</sub>	50.4 <sub>±11.7</sub>	31.1 <sub>±10.1</sub>	41.0 <sub>±10.0</sub>
	BoN	84.2 <sub>±0.0</sub>	51.0 <sub>±0.0</sub>	45.9 <sub>±0.0</sub>	72.2 <sub>±0.0</sub>	57.5 <sub>±0.0</sub>	62.1 <sub>±0.0</sub>	41.2 <sub>±0.0</sub>	51.0 <sub>±0.0</sub>
	Debate	82.7 <sub>±1.5</sub>	48.4 <sub>±2.6</sub>	43.4 <sub>±2.5</sub>	69.1 <sub>±3.1</sub>	60.4 <sub>±2.9</sub>	65.2 <sub>±3.1</sub>	42.9 <sub>±1.7</sub>	49.1 <sub>±1.9</sub>
	DyLAN	91.5 <sub>±7.3</sub>	63.1 <sub>±12.1</sub>	51.6 <sub>±5.7</sub>	80.4 <sub>±8.2</sub>	55.1 <sub>±2.4</sub>	58.0 <sub>±4.1</sub>	40.4 <sub>±0.8</sub>	45.5 <sub>±5.5</sub>
	GPTSwarm	91.5 <sub>±7.3</sub>	64.7 <sub>±13.7</sub>	52.4 <sub>±6.5</sub>	79.6 <sub>±7.4</sub>	56.8 <sub>±0.7</sub>	60.2 <sub>±1.9</sub>	40.3 <sub>±0.9</sub>	49.7 <sub>±1.3</sub>
	MacNet	83.8 <sub>±0.4</sub>	52.9 <sub>±1.9</sub>	46.2 <sub>±0.3</sub>	70.5 <sub>±1.7</sub>	61.8 <sub>±4.3</sub>	68.4 <sub>±6.3</sub>	46.4 <sub>±5.2</sub>	53.7 <sub>±2.7</sub>
	M2CL (ours)	95.1 <sub>±10.9</sub>	72.5 <sub>±21.5</sub>	78.9 <sub>±33.0</sub>	90.7 <sub>±18.5</sub>	79.0 <sub>±21.5</sub>	88.9 <sub>±26.8</sub>	67.2 <sub>±26.0</sub>	70.5 <sub>±19.5</sub>

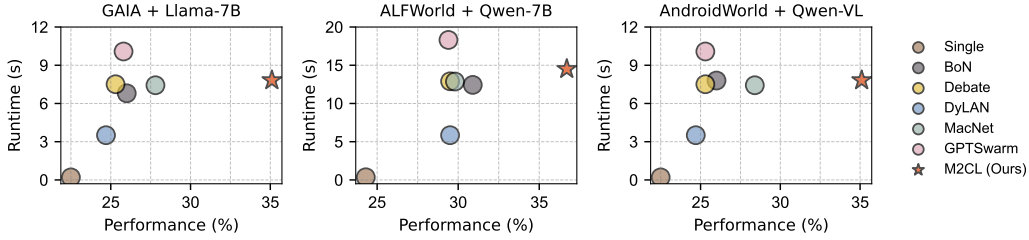


Figure 3: Performance versus runtime under different settings. Circles closer to the lower-left corner indicate higher efficiency.

## 6 EXPERIMENT

In this section, we conduct experiments to evaluate the performance of M2CL by answering the following research questions:

- Q1. How does M2CL perform compared to existing methods across various benchmarks, especially in complex agentic tasks?
- Q2. How does the performance of MAD scale with the number of LLMs?
- Q3. How is the performance affected by factors such as context constraint and components such as context initialization and context evolution?
- Q4. How do contexts promote consensus, and are they transferable to other models?

### 6.1 EXPERIMENTAL SETUP

**Dataset.** We run experiments with 3 domains including 9 datasets: **1) LLM reasoning**, including MMLU (Hendrycks et al., 2021), MATH (Hendrycks et al., 2021), and GPQA (Rein et al., 2023), HumanEval (Chen et al., 2021). **2) Embodied Agentic**, including ALFWorld Shridhar et al. (2021), SciWorld Wang et al. (2022), GAIA (Mialon et al., 2024), and PDDL Chang et al. (2024). **3) Mobile GUI** AndroidWorld Rawles et al. (2025). Details on datasets can be found in Appendix F.1.

**Baselines.** We evaluate our method against six strong baseline methods: **1) Single** execution, querying a single LLM to solve the task. **2) Best-of-N**, querying a single LLM N times and sampling the most correct answer. **3) Debate** (Du et al., 2023), a multi-agent framework where LLMs discuss their responses and reasoning processes over multiple rounds. **4) DyLAN** (Liu et al., 2024), a multi-agent framework where LLMs score each other and collaborate dynamically. **5) GPTSwarm** (Zhuge et al., 2024), a multi-agent framework that refines LLM prompts and improves



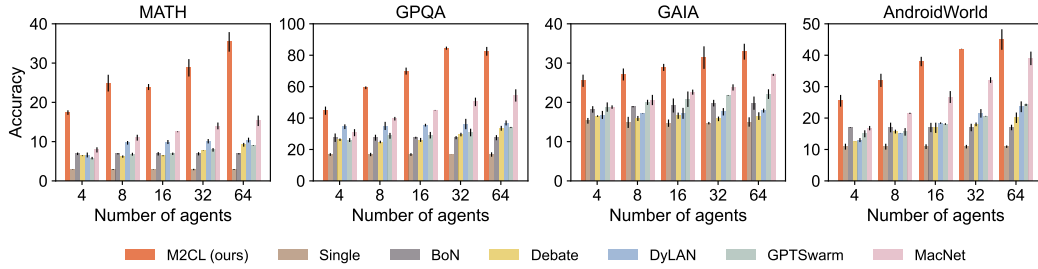


Figure 4: Performance of varying the numbers of LLMs. Uncertainty intervals depict standard deviation over three seeds.

LLM orchestration by changing their connectivity. **6) MacNet** (Jiang et al., 2023), a recent multi-LLM framework where LLMs are invoked between LLM interactions to provide actionable instructions to the next LLM based on the previous LLM’s outputs.

**Reproducibility.** All details of our experiments are provided in the appendices in terms of the tasks, network architectures, hyperparameters, etc. We conduct experiment on two series of LLM (Llama-2 (Touvron et al., 2023) and Qwen-2.5 (Yang et al., 2024a)) and three model sizes each for LLM reasoning, and Qwen2.5-VL (Bai et al., 2025) for GUI reasoning. All the experiments are run on Ubuntu 22.04.4 LTS with 8 NVIDIA H800 GPUs.

## 6.2 EXPERIMENTAL RESULTS

**Comparative results.** To answer the first question, we evaluate M2CL’s performance across all datasets, with varying base models and number of LLMs (ranging from 4 to 64). We select Qwen series models and 4 LLMs participating in Table 1 and provide full results in Appendices G.2 and G.3. We find M2CL consistently outperforms baselines in all 9 datasets, often by a significant margin in terms of performance. Of note, BoN outperforms most of the baselines especially in complex multi-round agentic tasks, revealing the drawback of fixed contexts, which, despite expanding the exploration space, do not converge and thus hinder LLMs from achieving true cooperative reasoning. In contrast, M2CL can adapt contexts and enhance the relevance of responses and questions while ensuring creativity, indicating that M2CL well avoids that LLMs with different contexts easily influence each other and successfully brings LLMs into cooperation by reaching a consensus through discussion.

In addition, we evaluate the efficiency of M2CL by visualizing the performance versus runtime. As shown in Fig. 3, M2CL consistently delivers the highest performance improvement (more than 20%) while maintaining a modest increase in runtime (less than 10%). This clearly demonstrates the efficiency of M2CL due to its lightweight context generator.

**Multi-agent scaling law.** To answer the second question, we run experiments with varying numbers of LLMs (ranging from 4 to 64). The data and parameter setup adhere to that of Table 2. We present selected results in Fig. 4 and full results in Tables 4 to 8 and Figs. 5 to 10 of Appendix G.2. Scaling our method reveals a more efficient scaling law as the performance grows logarithmically before saturation and improves faster than baselines. We speculate this arises because collaborative instructions in our generated contexts enable genuine inter-LLM cooperation, thereby unleashing the multidimensional reasoning capabilities of the MAD.

**Context constraint.** To answer the third question, we vary the context constraint  $\beta$  from 0 to 10 and run experiments across all datasets. Full results are shown in Figs. 12 to 16 of Appendix G.4. The results clearly indicate that as  $\beta$  increases, there is an initial improvement in performance; once it reaches a sufficiently large value, performance tends to drop. A strict (refers to a smaller  $\beta$ ) context constraint leads to generated contexts getting closer to the initial contexts, thereby causing discussion inconsistency. On the other hand, a loose context constraint leads to naive consistency, where LLMs tend to generate the same answers, resulting in insufficient creativity.

**Ablation studies.** We assess the effect of key components by ablating them on all datasets under the same setting. 1) *Importance of context initialization.* 2) *Importance of tuning  $\alpha$ .* 3) *Importance of context evolution.* As illustrated in Tables 11 to 13, LLMs struggle to specialize and coordinate

efficiently, whereas context initialization enables them to acquire high-impact contexts, significantly enhancing the foundational capabilities of MAD. Without tuning  $\alpha$  during discussion rounds, LLMs tend to reach an agreement in the first round, leading to responses that lack creativity and diversity, which ultimately reduces problem-solving ability. Without context evolution, LLMs lack collaborative guidance considering previous responses, which prevents them from effectively leveraging the outputs of other LLMs.

**Discrepancy intensity.** To answer the fourth question, we explored the change of Discrepancy intensity ( $\max_{i,j \in [N]} \|a_i - a_j\|^2$ ) with rounds to verify whether the generated context can help the LLMs gradually reach agreement with the discussion. Complete results are shown in Figs. 17 to 21, of Appendix G.5. We find that discrepancy intensity from M2CL decreases faster than other methods, corroborating its efficacy and superiority in converging the search space of multiple LLMs.

**Transferability of contexts.** To answer the fourth question, we also conduct experiments on transferring contexts directly into stronger LLMs to verify whether the generated contexts have better interpretation ability and efficacy. Full results are depicted in Table 10 of Appendix G.6. The results show that the transferred contexts deliver consistent improvement, indicating that the trained context generator can directly adapt to a wide range of models without additional retraining.

## 7 LIMITATION AND DISCUSSION

In this paper, we propose a novel context learning method, designed to expand LLMs’ horizons and reach consensus in MAD. By initializing and adjusting LLMs’ contexts based on the problem and discussion states, our method significantly improves problem-solving capabilities across diverse benchmarks while maintaining computational efficiency. A limitation of M2CL is the MAD framework in which diversity is brought about by the number of LLMs with heterogeneous characteristics which is computationally inefficient. An avenue for future work is to enable LLMs to truly capture the sub-tasks they are interested in or excel at.

## ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China under Grants 62572496 and 62402269, the Shenzhen Science and Technology Program under Grant JCYJ20250604175500001, and the Young Elite Scientist Sponsorship Program by CAST under Contract ZB2025-218.

## ETHICS STATEMENT

This work advances the field of large language model (LLM) research by introducing a context learning method for multi-agent discussion, demonstrating significant improvements in accuracy, diversity, and consensus across multiple benchmarks.

However, the broader implications of deploying such systems warrant careful consideration. Multi-agent discussion, while powerful, may inadvertently propagate biases introduced during context initialization or amplify errors during consensus-building. These risks are particularly critical in high-stakes domains like legal, financial, and healthcare applications, where erroneous outputs could lead to significant ethical, social, or economic consequences.

## REPRODUCIBILITY STATEMENT

The full algorithmic details of M2CL are presented in the main paper, with additional implementation details and hyperparameter settings included in Section 6. We provide source code at <https://github.com/HansenHua/M2CL-ICLR26>, which includes scripts for training, evaluation, and reproducing all reported experiments. We provide checkpoints at <https://huggingface.co/hansenhua/M2CL>. For theoretical results, we include complete proofs in Appendices B to D.

## REFERENCES

- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *Proceedings of The 11st International Conference on Learning Representations*, 2023.
- Ma Chang, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. In *Proceedings of Advances in Neural Information Processing Systems*, volume 37, pp. 74325–74362, 2024.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of The 40th International Conference on Machine Learning*, 2023.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. Healai: A healthcare llm for effective medical documentation. In *Proceedings of The 17th ACM International Conference on Web Search and Data Mining*, pp. 1167–1168, 2024.
- Qiuhan Gu. Llm-based code generation method for golang compiler testing. In *Proceedings of The 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 2201–2203, 2023.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of The 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023.
- Doohyuk Jang, Yoonjeon Kim, Chanjae Park, Hyun Ryu, and Eunho Yang. Reasoning model is stubborn: Diagnosing instruction overriding in reasoning models. *arXiv preprint arXiv:2505.17225*, 2025.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.
- Dongfang Li, Zhenyu Liu, Xinshuo Hu, Zetian Sun, Baotian Hu, and Min Zhang. In-context learning state vector with inner and momentum optimization. In *Proceedings of Advances in Neural Information Processing Systems*, 2024a.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Proceedings of Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023.
- Zhuowei Li, Zihao Xu, Ligong Han, Yunhe Gao, Song Wen, Di Liu, Hao Wang, and Dimitris N Metaxas. Implicit in-context learning. *arXiv preprint arXiv:2405.14660*, 2024b.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.
- Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration. In *Proceedings of The 1st Conference on Language Modeling*, 2024.
- Ziru Liu, Cheng Gong, Xinyu Fu, Yaofang Liu, Ran Chen, Shoubo Hu, Suiyun Zhang, Rui Liu, Qingfu Zhang, and Dandan Tu. Ghpo: Adaptive guidance for stable and efficient llm reinforcement learning. *arXiv preprint arXiv:2507.10628*, 2025.
- Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*, 2024.
- Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *Proceedings of The 11th International Conference on Learning Representations*, 2023.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Proceedings of Advances in Neural Information Processing Systems*, volume 36, 2024.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *Proceedings of The 12th International Conference on Learning Representations*, 2024.
- Deepak Pandita, Tharindu Cyril Weerasooriya, Ankit Parag Shah, Isabelle Diana May-Xin Ng, Christopher M Homan, and Wei Wei. Prorrefine: Inference-time prompt refinement with textual feedback. *arXiv preprint arXiv:2506.05305*, 2025.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of The 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.
- Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*, 2023.
- Kiran Purohit, V Venkatesh, Sourangshu Bhattacharya, and Avishek Anand. Sample efficient demonstration selection for in-context learning. In *Proceedings of The 42nd International Conference on Machine Learning*, 2025.

- Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. Scaling large language model-based multi-agent collaboration. In *Proceedings of The 13th International Conference on Learning Representations*, 2025.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *Proceedings of The 12th International Conference on Learning Representations*, 2023.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. Planning with large language models via corrective re-prompting. In *Proceedings of The 36th Conference on Neural Information Processing System Foundation Models for Decision Making Workshop*, 2022.
- Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. In *Proceedings of The Thirteenth International Conference on Learning Representations*, 2025.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.
- Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. Small llms are weak tool learners: A multi-llm agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16658–16680, 2024.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *Proceedings of The 10th International Conference on Learning Representations*, 2021.
- Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. In *Proceedings of The 41st International Conference on Machine Learning*, 2024.
- Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. Chemreasoner: Heuristic search over a large language model’s knowledge space using quantum-chemical feedback. In *Proceedings of The 41st International Conference on Machine Learning*, 2024.
- Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *Proceedings of The 12th International Conference on Learning Representations*, 2024.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- A Vaswani. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, 2017.

- Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of The 41st International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.
- Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022.
- Sizhe Wang, Zhengren Wang, Dongsheng Ma, Yongan Yu, Rui Ling, Zhiyu Li, Feiyu Xiong, and Wentao Zhang. Codeflowbench: A multi-turn, iterative benchmark for complex code generation. *arXiv preprint arXiv:2504.21751*, 2025.
- Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*, 2023.
- Shiwei Wu, Yu Liu, Yan Gao, Zhi Li, Chao Zhang, Qimeng Wang, Enhong Chen, and You He. Evaluating conversational question answering in multi-instructional documents with large language models. *Chinese Journal of Electronics*, 2025.
- Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng YANG, Qingxing Cao, Haiming Wang, Xiongwei Han, Jing Tang, Chengming Li, and Xiaodan Liang. DQ-lore: Dual queries with low rank approximation re-ranking for in-context learning. In *Proceedings of The 12th International Conference on Learning Representations*, 2024.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.
- Hao Yang, Qianghua Zhao, and Lei Li. Chain-of-thought in large language models: Decoding, projection, and activation. *arXiv preprint arXiv:2412.03944*, 2024b.
- Jingyuan Yang, Rongjun Li, Weixuan Wang, Ziyu Zhou, Zhiyong Feng, and Wei Peng. Lf-steering: Latent feature activation steering for enhancing semantic consistency in large language models. *arXiv preprint arXiv:2501.11036*, 2025.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *Proceedings of The 11th International Conference on Learning Representations*, 2023.
- Beiyuan Yu, Jian Mao, and Jianwei Liu. Cookiechecker: Automated cookie setting compliance analysis in real-world web environments. *Chinese Journal of Electronics*, 2025.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022.
- Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*, 2025a.
- Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. In *Proceedings of The 13th International Conference on Learning Representations*, 2025b.
- Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In *Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9134–9148, 2022.
- Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*, 2024.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zihan Wang, Andi Wang, Yang Li, et al. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5673–5684, 2023.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Gptswarm: Language agents as optimizable graphs. In *Proceedings of The 41st International Conference on Machine Learning*, 2024.

## A USAGE OF LLMs

We use large language models (LLMs) solely as an assistive tool for polishing the writing and improving clarity of exposition. All content generated by LLMs was carefully reviewed, verified, and, where necessary, revised by the authors. The authors take full responsibility for the correctness and integrity of the final manuscript.

## B PROOF OF THEOREM 4.1

*Proof.* Define the optimal  $\omega^* = \arg \min_{\omega} \|a_c - \sum_{i=1}^N \omega_i a(C_i^b)\|$  and its corresponding activation  $a = \sum_{i=1}^N \omega_i^* a(C_i^b)$ . Then, we can derive the upper bound of the activation difference using the triangle inequality:

$$\sum_{i=1}^N \|a_c - a(C_i^t)\| \leq \sum_{i=1}^N \|a(C_i^t) - a\| + N\|a - a_c\|. \quad (18)$$

Using the triangle inequality, we can bound the first term in Eq. (18):

$$\begin{aligned} \sum_{i=1}^N \|a(C_i^t) - a\| &= \sum_{i=1}^N \left\| \sum_{j=1}^N \omega_j^* (a(C_i^t) - a(C_j^b)) \right\| \\ &\leq \sum_{i=1}^N \sum_{j=1}^N |\omega_j^*| \cdot \|a(C_i^t) - a(C_j^b)\| \leq \sum_{i=1}^N \sum_{j=1}^N \|a(C_i^t) - a(C_j^b)\| \\ &\leq \sum_{i=1}^N \sum_{j=1}^N \|a(C_i^t) - a(C_j^t)\| + \sum_{i=1}^N \sum_{j=1}^N \|a(C_j^t) - a(C_j^b)\| \\ &\leq \sum_{i=1}^N \sum_{j=1}^N \|a(C_i^t) - a(C_j^t)\| + \sum_{i=1}^N \sum_{j=1}^N L_a \|C_j^t - C_j^b\|. \end{aligned} \quad (19)$$

Then, we bound the second term in Eq. (18):

$$\begin{aligned} \|a - a_c\| &\leq \left\| \sum_{i=1}^N \omega_i^* a(C_i^t) - a_c \right\| = \left\| \sum_{i=1}^N \omega_i^* a(C_i^b) + \sum_{i=1}^N \omega_i^* (a(C_i^t) - a(C_i^b)) - a_c \right\| \\ &\leq \min_{\omega} \left\| a_c - \sum_{i=1}^N \omega_i a(C_i^b) \right\| + \sum_{i=1}^N \|\omega_i^* (a(C_i^t) - a(C_i^b))\| \\ &\leq \min_{\omega} \left\| a_c - \sum_{i=1}^N \omega_i a(C_i^b) \right\| + \sum_{i=1}^N \|a(C_i^t) - a(C_i^b)\| \\ &\leq \min_{\omega} \left\| a_c - \sum_{i=1}^N \omega_i a(C_i^b) \right\| + L_a \sum_{i=1}^N \|C_i^t - C_i^b\|. \end{aligned} \quad (20)$$

In the first line, we add and subtract  $\sum_{i=1}^N \omega_i^* a(C_i^b)$ . In the second line, we use the triangle inequality and the definition of  $\omega^*$ . In the third line, we scale  $\omega^*$  to 1. Finally, we utilize the smoothness property of the activation function to bound the second term. Substituting into Eq. (18), we can derive:

$$\begin{aligned} \sum_{i=1}^N \|a_c - a(C_i^t)\| &\leq \sum_{i=1}^N \left( \sum_{j=1}^N \|a(C_i^t) - a(C_j^t)\| + (N+1)L_a \|C_i^t - C_i^b\| \right) \\ &\quad + N \min_{\omega} \left\| a_c - \sum_{i=1}^N \omega_i a(C_i^b) \right\|, \end{aligned} \quad (21)$$

□



## C DECOUPLED CRITERION FUNCTION

**Lemma C.1.** *Under the assumption of one-block transformer, LLMs' activation diversity can be bound by the activation difference between their instructions and responses.*

$$\|a(C_i^t) - a(C_j^t)\| \leq \|a(I_i^t) - a(X_i^t)\| + \|a(I_j^t) - a(X_j^t)\| + 18L_V n N \exp(2\rho^2). \quad (22)$$

*Proof.* We first decompose the activation of context into a combination of its components:

$$\begin{aligned} a(C_i^t) &= a(C_i^t) - [a(I_i^t) + \sum_{k \neq i}^N a(X_k^{t-1}) - (N-1)a(P)] + [a(I_i^t) + \sum_{k \neq i}^N a(X_k^{t-1}) - (N-1)a(P)] \\ a(C_j^t) &= a(C_j^t) - [a(I_j^t) + \sum_{k \neq j}^N a(X_k^{t-1}) - (N-1)a(P)] + [a(I_j^t) + \sum_{k \neq j}^N a(X_k^{t-1}) - (N-1)a(P)]. \end{aligned} \quad (23)$$

By using Lemma D.3, we can derive the upper bound of the activation difference as:

$$\begin{aligned} &\|a(C_i^t) - a(C_j^t)\| \\ &\leq \|a(C_i^t) - [a(I_i^t) + \sum_{k \neq i}^N a(X_k^{t-1}) - (N-1)a(P)]\| + \|a(I_i^t) - a(X_i^t)\| \\ &\quad + \|a(C_j^t) - [a(I_j^t) + \sum_{k \neq j}^N a(X_k^{t-1}) - (N-1)a(P)]\| + \|a(I_j^t) - a(X_j^t)\| \\ &\leq \|a(I_i^t) - a(X_i^t)\| + \|a(I_j^t) - a(X_j^t)\| + 18L_V n N \exp(2\rho^2). \end{aligned} \quad (24)$$

□

## D USEFUL LEMMA

**Lemma D.1.** *Suppose  $\|W_V X\| \leq L_V$  and  $\|(W_K X)^T W_Q X\| \leq \rho^2$ , the difference between the activation of softmax attention  $a(X)$  and linear attention  $a'(X)$  can be bounded by:*

$$\|a(X) - a'(X)\| \leq 3L_V n \exp(2\rho^2) \quad (25)$$

where  $a(X) = W_V X \text{softmax}(\frac{(W_K X)^T W_Q X}{\sqrt{d}})$  and  $a'(X) = W_V X (W_K X)^T W_Q X$ .

*Proof.* Define  $\Delta = D^{-1} \exp(S) - \exp(S)$ , where  $S = (W_K X)^T W_Q X$  and  $D = \text{diag}(\langle \exp(S), \mathbf{1}_n \rangle)$ . Then, we can derive the upper bound of each element of  $\Delta$ .

$$\Delta_{ij} = \frac{\exp(S_{ij})}{D_{ii}} - \exp(S_{ij}) \leq \exp(S_{ij}) \left( \left| \frac{1}{D_{ii}} \right| + 1 \right) \leq \exp(\rho^2) (1 + \exp(\rho^2)) \leq 2 \exp(2\rho^2) \quad (26)$$

Summing them up, we can derive:

$$\|D^{-1} \exp(S) - \exp(S)\| \leq \|D^{-1} \exp(S) - \exp(S)\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n \Delta_{ij}^2} \leq 2n \exp(2\rho^2) \quad (27)$$

Define  $\delta = \exp(S) - S$ , where  $S = (W_K X)^T W_Q X$ . Then, we derive the upper bound of  $\delta$  by finding the upper bound of each element of  $\delta$ .

$$\delta_{ij} = \exp(S_{ij}) - S_{ij} \leq \max\{\exp(\rho^2) - \rho^2, \rho^2 + \exp(-\rho^2)\} \leq \exp(\rho^2) \quad (28)$$

Summing them up, we can derive:

$$\|\exp(S) - S\| \leq \|\exp(S) - S\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n \exp(S_{ij}) - S_{ij}} \leq n \exp(\rho^2) \quad (29)$$

After bounding  $\|D^{-1} \exp(S) - \exp(S)\|$  and  $\|\exp(S) - S\|$ , we can derive the upper bound of  $\|D^{-1} \exp(S) - S\|$  by basic algebra as:

$$\begin{aligned} & \|D^{-1} \exp(S) - S\| \\ &= \|D^{-1} \exp(S) - \exp(S) + \exp(S) - S\| \\ &\leq \|D^{-1} \exp(S) - \exp(S)\| + \|\exp(S) - S\| \\ &\leq 2n \exp(2\rho^2) + n \exp(\rho^2) \\ &\leq 3n \exp(2\rho^2) \end{aligned} \quad (30)$$

Therefore, we derive the difference between the activation of softmax attention and linear attention as:

$$\begin{aligned} & \|W_V X \text{softmax}\left(\frac{(W_K X)^T W_Q X}{\sqrt{d}}\right) - W_V X (W_K X)^T W_Q X\| \\ &\leq \|W_V X\| \cdot \|\text{softmax}\left(\frac{(W_K X)^T W_Q X}{\sqrt{d}}\right) - (W_K X)^T W_Q X\| \\ &\leq L_V \cdot \|D^{-1} \exp(S) - S\| \\ &\leq 3L_V n \exp(2\rho^2) \end{aligned} \quad (31)$$

□

**Lemma D.2.** Define the activation of linear attention:

$$\begin{aligned} a'(Y) &\doteq W_V[Y, P](W_K[Y, P])^T W_Q Y \\ a'(P) &\doteq W_V P (W_K P)^T W_Q P, \end{aligned} \quad (32)$$

then the activation of a long prompt  $Y = [Y_1, Y_2, \dots, Y_N]$  can be derived by the activation of its component:

$$a'(Y) = \sum_{i=1}^N a'(Y_i) - (N-1)a'(P). \quad (33)$$

*Proof.*

$$\begin{aligned} a'(Y) &= W_V[Y_1, Y_2, \dots, Y_N, P] \left( W_K[Y_1, Y_2, \dots, Y_N, P] \right)^T W_Q P \\ &= W_V[Y_1, Y_2, \dots, Y_N, P] [Y_1, Y_2, \dots, Y_N, P]^T W_K^T W_Q P \\ &= W_V \left( \sum_{i=1}^N Y_i Y_i^T + P P^T \right) W_K^T W_Q P \\ &= W_V \left[ \sum_{i=1}^N (Y_i Y_i^T + P P^T) - (N-1) P P^T \right] W_K^T W_Q P \\ &= \sum_{i=1}^N W_V (Y_i Y_i^T + P P^T) W_K^T W_Q P - W_V (N-1) P P^T W_K^T W_Q P \\ &= \sum_{i=1}^N W_V [Y_i, P] (W_K [Y_i, P])^T W_Q P - (N-1) W_K P (W_K P)^T W_Q P \\ &= \sum_{i=1}^N a'(Y_i) - (N-1)a'(P). \end{aligned} \quad (34)$$

□

**Lemma D.3.** *We derive the difference between the activation of a long prompt  $Y$  and a combination of the activations of its components  $Y = [Y_1, Y_2, \dots, Y_N, P]$  as:*

$$\|a(Y) - \sum_{i=1}^N a(Y_i) + (N-1)a(P)\| \leq 9L_V n N \exp(\rho^2). \quad (35)$$

*Proof.* Denote the activation of linear attention as  $a'$ , we can complete the proof by using Lemma D.2 to simplify the relationship of linear attention and Lemma D.1 to bound the difference:

$$\begin{aligned} & \|a(Y) - \sum_{i=1}^N a(Y_i) + (N-1)a(P)\| \\ &= \|a(Y) - a'(Y) + a'(Y) - [\sum_{i=1}^N a'(Y_i) - (N-1)a'(P)] \\ & \quad + [\sum_{i=1}^N a'(Y_i) - (N-1)a'(P)] - [\sum_{i=1}^N a(Y_i) - (N-1)a(P)]\| \\ &\leq \|a(Y) - a'(Y)\| + \sum_{i=1}^N \|a'(Y_i) - a(Y_i)\| + (N-1)\|a'(P) - a(P)\| \\ &\leq \exp(2\rho^2)[3L_V n N + 3L_V N n + 3L_V (N-1)n] \\ &\leq 9L_V n N \exp(2\rho^2) \end{aligned} \quad (36)$$

□

**Lemma D.4.** *If  $\|S\| \leq \rho$ , we have  $\langle \exp(S), \mathbf{1}_n \rangle \geq \exp(-\rho)$*

*Proof.*

$$\begin{aligned} \langle \exp(S), \mathbf{1}_n \rangle &= \sum_{i=1}^n \exp(S_i) \geq \min_{i \in [n]} \exp(S_i) \geq \min_{i \in [n]} \exp(-|S_i|) \\ &= \exp(-\max_{i \in [n]} |S_i|) = \exp(-\|S\|_\infty) \geq \exp(-\|S\|_2) \geq \exp(-\rho) \end{aligned} \quad (37)$$

□

## E MULTI-ROUND CONTEXT LEARNING

First, we define:

$$h(I_i^t) \doteq \beta - \|I_i^t - I_i^b\| \quad (38)$$

$$f(I_i^t) \doteq \begin{cases} \|a(I_i^t) - a(X_i^{t-1})\|, & \text{if } h(I_i^t) \geq 0 \\ +\infty, & \text{otherwise.} \end{cases} \quad (39)$$

To solve the minimization optimization with inequality constraint, we can construct a Lagrangian expression with a Lagrange multiplier  $\alpha_T$ :

$$\text{minimize } f(I_i^t) \text{ s.t. } h(I_i^t) \geq 0 \Leftrightarrow L(I_i^t, \alpha_t) = f(I_i^t) - \alpha_t h(I_i^t). \quad (40)$$

The optimization changes to:

$$\min_{I_i^t} f(I_i^t) = \max_{\alpha_t \geq 0} \min_{I_i^t} L(I_i^t, \alpha_t). \quad (41)$$

Therefore, to minimize  $f(I_i^t)$ , the dual problem is listed as below. Note that to make sure  $\min_{I_i^t} f(I_i^t)$  is properly minimized and would not become  $+\infty$ , the constraint has to be satisfied.

$$\begin{aligned}
& \min_{I_i^t} \left[ \|a(I_i^t) - a(X_i^{t-1})\| \right] \\
&= \min_{I_i^t} f(I_i^t) \\
&= \max_{\alpha_t \geq 0} \min_{I_i^t} L(I_i^t, \alpha_t) \\
&= \max_{\alpha_t \geq 0} \min_{I_i^t} f(I_i^t) - \alpha_t h(I_i^t) \\
&= \max_{\alpha_T \geq 0} \min_{I_i^t} \left[ \|a(I_i^t) - a(X_i^{t-1})\| \right] - \alpha_t (\beta - \|I_i^t - I_i^b\|) \\
&= \max_{\alpha_T \geq 0} \min_{I_i^t} \left[ \|a(I_i^t) - a(X_i^{t-1})\| - \alpha_T \beta + \alpha_t \|I_i^t - I_i^b\| \right]. \tag{42}
\end{aligned}$$

Define the Q-function for this multi-round multi-agent collaboration as:

$$Q(a(I_i^t), a(X_i^{t-1})) = \|a(I_i^t) - a(X_i^{t-1})\| + \min_{I_i^t} \|a(I_i^{t+1}) - a(X_i^t)\| + \alpha_{t+1} \|I_i^{t+1,*} - I_i^b\|. \tag{43}$$

Here,  $I_i^{t+1,*}$  denotes the optimal  $I_i^{t+1}$ . Therefore, the expected return is as follows, when we take one round further back to the round  $T-1$ :

$$\begin{aligned}
& \min_{I_i^{T-1}} \left( \|a(I_i^{T-1}) - a(X_i^{T-2})\| + \min_{I_i^t} \|a(I_i^t) - a(X_i^{T-1})\| \right) \\
&= \min_{I_i^{T-1}} \left( Q(a(I_i^{T-1}), a(X_i^{T-2})) - \alpha_T^* \|I_i^{T,*} - I_i^b\| \right) \\
&= \max_{\alpha_{T-1} > 0} \min_{I_i^{T-1}} \left( Q(a(I_i^{T-1}), a(X_i^{T-2})) - \alpha_{T-1} (\beta - \|I_i^{T-1} - I_i^b\|) - \alpha_T^* \|I_i^{T,*} - I_i^b\| \right) \\
&= \max_{\alpha_{T-1} > 0} \min_{I_i^{T-1}} \left( Q(a(I_i^{T-1}), a(X_i^{T-2})) - \alpha_{T-1} \beta + \alpha_{T-1} \|I_i^{T-1} - I_i^b\| \right) - \alpha_T^* \|I_i^{T,*} - I_i^b\|. \tag{44}
\end{aligned}$$

Similar to the previous round,

$$\begin{aligned}
I_i^t &= \arg \min_{I_i^t} \left[ \|a(I_i^t) - a(X_i^{t-1})\| - \alpha_t \beta + \alpha_t \|I_i^t - I_i^b\| \right] \\
\alpha_t^* &= \arg \max_{\alpha_t \geq 0} \left[ \alpha_T \|I_i^t - I_i^b\| - \alpha_t \beta \right]. \tag{45}
\end{aligned}$$

By repeating this process, we can learn the optimal temperature parameter in every round. Hence, the loss functions for  $\theta$  and  $\alpha$  are as follows:

$$\begin{aligned}
L(\theta_i^t) &= \|a(I_i^t) - a(X_i^{t-1})\| + \alpha_t \|I_i^t - I_i^b\| \\
L(\alpha) &= \alpha (\beta - \|I_i^t - I_i^b\|). \tag{46}
\end{aligned}$$

## F EXPERIMENT SETUP

### F.1 DATASETS

We evaluate our method on three areas with 7 datasets which are widely used in prior studies (Dubey et al., 2024; Qian et al., 2025). We elaborate on what follows.

- MMLU (Hendrycks et al., 2021), a comprehensive benchmark covering diverse subjects and difficulty levels, designed to test world knowledge and logical reasoning through multiple-choice questions.
- MATH (Hendrycks et al., 2021), a dataset of challenging competition-level math problems requiring multi-step symbolic reasoning and advanced problem-solving ability.
- GPQA (Rein et al., 2023), a benchmark of graduate-level multiple-choice science questions that assesses deep domain knowledge and reasoning under uncertainty.
- HumanEval (Chen et al., 2021), a widely recognized benchmark for function-level code generation, designed to evaluate fundamental programming skills.
- ALFWorld (Shridhar et al., 2021), a text-based embodied environment featuring household tasks, where agents navigate and interact with objects via natural language commands.
- SciWorld (Wang et al., 2022), a text-based embodied environment for interactive science tasks, requiring agents to navigate rooms, conduct experiments, and perform procedural reasoning.
- GAIA (Mialon et al., 2024), a benchmark of real-world question answering tasks that integrate knowledge retrieval, reasoning, and multi-step tool use.
- PDDL (Chang et al., 2024), an environment comprising diverse strategic games, where agents must employ PDDL expressions to plan and execute complex tasks.
- AndroidWorld (Rawles et al., 2025), an environment with 116 dynamic tasks across 20 real-world Android apps, designed to evaluate mobile agents’ capabilities in app navigation and system-level control.

We use 20% of the questions to construct the training dataset and the rest as testing dataset.

### F.2 BASELINES

We test our method against six baselines. We implement them based on their publicly available implementations.

- *Single Execution* (`Single`), querying a single LLM to solve the task.
- *Best-of-N sampling* (`BON`), querying a single LLM  $N$  times and sampling the most correct answer. We set  $N$  to 32, as increasing  $N$  further does not yield additional performance benefits and aligns closely with the number of LLM calls used for multi-agent methods.
- *Discussion* (Du et al., 2023), a multi-agent framework in which LLMs are assigned predefined distinct contexts and iteratively exchange their reasoning processes as additional prompts over multiple rounds, before producing a final answer via majority voting.
- *Dynamic LLM-Powered Agent Network* (Liu et al., 2024) (`DYLAN`), a discussion-style framework which incorporates an LLM selection algorithm based on an unsupervised metric, namely the Agent Importance Score, which identifies the most contributive LLMs through a preliminary trial tailored to the specific task.
- *GPTSwarm* (Zhuge et al., 2024), formalizing a swarm of autonomous agents as computational graphs, with nodes as manually-customized functions and edges facilitating information flow, adaptively optimizing node prompts and modifying graph connectivity during collective reasoning.
- *Multi-LLM Collaboration Network* (Qian et al., 2025) (`MacNet`), a representative framework for decentralized and scalable multi-LLM systems. It introduces edge agents that mediate interactions by generating actionable instructions for the next agent based on the outputs of the previous one.

### F.3 IMPLEMENTATION DETAILS

The context pool is constructed using GPT-4o, where we prompt it to generate a large collection of high-quality initial contexts across diverse domains, including mathematics, science, coding, and

embodied reasoning, and various domain-specific sub-contexts are included in each domain. This ensures that the pool provides a broad coverage of reasoning perspectives and is shared across all tasks.

The context generators are implemented by T5-small model (Raffel et al., 2020) for all the tasks. The dimension of the generated context vectors is 512. The learning rates for the context generators and  $\alpha$  are both  $1e-4$ . We use 20% of the questions to train the context initialization and generator.

Table 2: Hyperparameters (identical across datasets).

Hyperparameter	Value
Sentence vector dimension	512
Optimizer	Adam
Batchsize	32
Learning rate of context	$1e-4$
Learning rate of $\alpha$	$1e-4$
Maximum rounds for discussion	8
Training rounds	100
Size of context pool	100

We implement our code using Pytorch 2.3.0, built upon the open-source parameters of the llama2, Qwen2.5, and Qwen2.5-VL. Models provided at <https://huggingface.co/meta-llama> and <https://huggingface.co/Qwen>. All the experiments are run on Ubuntu 22.04.4 LTS with 8 NVIDIA H800 GPUs.

#### F.4 PSEUDOCODE OF M2CL

We present the pseudocode of training M2CL in Algorithm 1. It begins with training the context initialization, then the context generators are trained along with the weight  $\alpha$  during the discussion.

**Algorithm 1:** Pseudocode of M2CL

---

```

1 Initialize parameters  $\phi_f, \phi_F, \{\theta_i\}_{i=1}^N$ , and  $\{\alpha_i\}_{i=1}^N$ ;
2 for each question do
3    $\phi_f \leftarrow \phi_f - \eta_f \nabla L(\phi_f), \quad \phi_F \leftarrow \phi_F - \eta_F \nabla L(\phi_F)$ ;
4 end
5 for each epoch do
6   Obtain initial contexts  $\{I_i^b\}_{i=1}^N$  via Eq. (12);
7   for each round  $t$  do
8     for  $i = 1$  to  $N$  do
9       Generate instructions  $I_i^t$  via Eq. (3) and obtain responses  $X_i^t$  via Eq. (4);
10       $\theta_i \leftarrow \theta_i - \eta_\theta \nabla L(\theta_i), \quad \alpha_i \leftarrow \alpha_i - \eta_\alpha \nabla L(\alpha_i)$ ;
11    end
12  end
13 end

```

---

## G ADDITIONAL RESULTS

### G.1 COMPARISON WITH MORE MODELS

To further investigate M2CL’s performance, we compare it with `best-of-N` using stronger close-sourced LLMs, including Qwen2.5-Max and GPT-4. As shown in Table 3, llama series models perform worse when fewer LLMs participate but achieve higher accuracy as the number of LLMs increases, demonstrating that M2CL can collaborate weaker models to achieve comparable performance.

Table 3: Accuracy using different base models on different datasets. We exhibit the performance advantage with Qwen-max and highlight the `best` result.

	Model	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
n=1	Qwen-max	86.5 <sup>±0.0</sup>	43.6 <sup>±0.0</sup>	35.7 <sup>±0.0</sup>	65.9 <sup>±0.0</sup>	73.1 <sup>±0.0</sup>	68.7 <sup>±0.0</sup>	61.5 <sup>±0.0</sup>	76.4 <sup>±0.0</sup>
	GPT-4	84.3 <sup>±2.2</sup>	41.1 <sup>±2.5</sup>	33.3 <sup>±2.4</sup>	62.7 <sup>±3.2</sup>	70.2 <sup>±2.9</sup>	65.4 <sup>±3.3</sup>	58.7 <sup>±2.8</sup>	73.5 <sup>±2.9</sup>
	Llama-7B	45.3 <sup>±41.2</sup>	2.9 <sup>±40.7</sup>	16.8 <sup>±18.9</sup>	15.8 <sup>±50.1</sup>	22.5 <sup>±50.6</sup>	20.1 <sup>±48.6</sup>	15.3 <sup>±46.2</sup>	24.2 <sup>±52.2</sup>
	Llama-13B	54.8 <sup>±31.7</sup>	3.9 <sup>±39.7</sup>	25.0 <sup>±10.7</sup>	23.7 <sup>±42.2</sup>	28.5 <sup>±44.6</sup>	24.5 <sup>±44.2</sup>	18.6 <sup>±42.9</sup>	30.4 <sup>±46.0</sup>
	Llama-70B	68.9 <sup>±17.6</sup>	31.6 <sup>±12.0</sup>	27.6 <sup>±8.1</sup>	35.5 <sup>±30.4</sup>	46.1 <sup>±27.0</sup>	40.2 <sup>±28.5</sup>	29.9 <sup>±31.6</sup>	48.4 <sup>±28.0</sup>
n=4	Qwen-max	87.5 <sup>±0.0</sup>	44.2 <sup>±0.0</sup>	39.2 <sup>±0.0</sup>	67.5 <sup>±0.0</sup>	74.0 <sup>±0.0</sup>	69.8 <sup>±0.0</sup>	62.5 <sup>±0.0</sup>	77.5 <sup>±0.0</sup>
	GPT-4	86.4 <sup>±1.1</sup>	43.9 <sup>±0.3</sup>	38.8 <sup>±0.4</sup>	67.0 <sup>±0.5</sup>	72.8 <sup>±1.2</sup>	68.0 <sup>±1.8</sup>	61.0 <sup>±1.5</sup>	76.0 <sup>±1.5</sup>
	Llama-7B	59.1 <sup>±28.4</sup>	17.4 <sup>±26.8</sup>	44.7 <sup>±5.5</sup>	32.9 <sup>±34.6</sup>	35.1 <sup>±38.9</sup>	33.0 <sup>±36.8</sup>	25.5 <sup>±37.0</sup>	33.1 <sup>±44.4</sup>
	Llama-13B	86.9 <sup>±0.6</sup>	23.4 <sup>±20.8</sup>	58.9 <sup>±19.7</sup>	47.5 <sup>±20.0</sup>	42.9 <sup>±31.1</sup>	39.9 <sup>±29.9</sup>	30.5 <sup>±32.0</sup>	40.9 <sup>±36.6</sup>
	Llama-70B	95.6 <sup>±8.1</sup>	40.3 <sup>±3.9</sup>	78.7 <sup>±39.5</sup>	70.6 <sup>±3.1</sup>	69.5 <sup>±4.5</sup>	65.5 <sup>±4.3</sup>	49.6 <sup>±13.0</sup>	66.5 <sup>±11.0</sup>
n=8	Qwen-max	88.5 <sup>±0.0</sup>	45.0 <sup>±0.0</sup>	40.5 <sup>±0.0</sup>	68.8 <sup>±0.0</sup>	75.0 <sup>±0.0</sup>	71.0 <sup>±0.0</sup>	63.5 <sup>±0.0</sup>	79.0 <sup>±0.0</sup>
	GPT-4	86.5 <sup>±2.0</sup>	44.6 <sup>±0.4</sup>	38.9 <sup>±1.6</sup>	67.1 <sup>±1.7</sup>	73.2 <sup>±1.8</sup>	69.5 <sup>±1.5</sup>	61.0 <sup>±2.5</sup>	76.0 <sup>±3.0</sup>
	Llama-7B	63.8 <sup>±24.7</sup>	24.9 <sup>±20.1</sup>	59.3 <sup>±18.8</sup>	38.8 <sup>±30.0</sup>	37.2 <sup>±37.8</sup>	36.1 <sup>±34.9</sup>	27.1 <sup>±36.4</sup>	34.7 <sup>±44.3</sup>
	Llama-13B	92.7 <sup>±4.2</sup>	23.0 <sup>±22.0</sup>	70.7 <sup>±30.2</sup>	55.4 <sup>±13.4</sup>	46.1 <sup>±28.9</sup>	44.7 <sup>±26.3</sup>	33.6 <sup>±29.9</sup>	43.8 <sup>±35.2</sup>
	Llama-70B	95.4 <sup>±6.9</sup>	43.5 <sup>±1.5</sup>	87.6 <sup>±47.1</sup>	81.0 <sup>±12.2</sup>	78.9 <sup>±3.9</sup>	76.2 <sup>±5.2</sup>	58.7 <sup>±4.8</sup>	73.6 <sup>±5.4</sup>
n=16	Qwen-max	88.5 <sup>±0.0</sup>	45.0 <sup>±0.0</sup>	40.5 <sup>±0.0</sup>	68.8 <sup>±0.0</sup>	75.0 <sup>±0.0</sup>	71.0 <sup>±0.0</sup>	63.5 <sup>±0.0</sup>	79.0 <sup>±0.0</sup>
	GPT-4	87.9 <sup>±0.6</sup>	44.8 <sup>±0.2</sup>	40.1 <sup>±0.4</sup>	68.5 <sup>±0.3</sup>	74.3 <sup>±0.7</sup>	70.5 <sup>±0.5</sup>	63.0 <sup>±0.5</sup>	78.2 <sup>±0.8</sup>
	Llama-7B	71.5 <sup>±17.0</sup>	23.9 <sup>±21.1</sup>	69.9 <sup>±29.4</sup>	48.3 <sup>±20.5</sup>	39.8 <sup>±35.2</sup>	38.4 <sup>±32.6</sup>	28.9 <sup>±34.6</sup>	37.0 <sup>±42.0</sup>
	Llama-13B	94.5 <sup>±6.0</sup>	32.7 <sup>±12.3</sup>	88.5 <sup>±48.0</sup>	66.1 <sup>±2.7</sup>	48.7 <sup>±26.3</sup>	47.1 <sup>±23.9</sup>	36.2 <sup>±27.3</sup>	45.5 <sup>±33.5</sup>
	Llama-70B	93.9 <sup>±5.4</sup>	51.5 <sup>±6.5</sup>	91.3 <sup>±50.8</sup>	97.2 <sup>±28.4</sup>	84.9 <sup>±9.9</sup>	81.7 <sup>±10.7</sup>	61.5 <sup>±2.0</sup>	78.1 <sup>±0.9</sup>
n=32	Qwen-max	89.0 <sup>±0.0</sup>	45.5 <sup>±0.0</sup>	41.0 <sup>±0.0</sup>	69.5 <sup>±0.0</sup>	75.5 <sup>±0.0</sup>	71.5 <sup>±0.0</sup>	64.0 <sup>±0.0</sup>	79.5 <sup>±0.0</sup>
	GPT-4	89.1 <sup>±0.1</sup>	45.4 <sup>±0.1</sup>	41.5 <sup>±0.5</sup>	69.7 <sup>±0.2</sup>	75.0 <sup>±0.5</sup>	71.0 <sup>±0.5</sup>	63.5 <sup>±0.5</sup>	78.5 <sup>±1.0</sup>
	Llama-7B	79.1 <sup>±9.9</sup>	28.8 <sup>±16.7</sup>	84.5 <sup>±43.5</sup>	56.5 <sup>±13.0</sup>	42.5 <sup>±33.0</sup>	40.7 <sup>±30.8</sup>	31.4 <sup>±32.6</sup>	38.9 <sup>±40.6</sup>
	Llama-13B	95.8 <sup>±6.8</sup>	41.4 <sup>±4.1</sup>	94.7 <sup>±53.7</sup>	75.6 <sup>±6.1</sup>	52.2 <sup>±23.3</sup>	50.7 <sup>±20.8</sup>	38.5 <sup>±25.5</sup>	48.0 <sup>±31.5</sup>
	Llama-70B	97.0 <sup>±8.0</sup>	54.5 <sup>±9.0</sup>	95.1 <sup>±54.1</sup>	93.7 <sup>±24.2</sup>	88.5 <sup>±13.0</sup>	86.0 <sup>±14.5</sup>	65.7 <sup>±1.7</sup>	79.8 <sup>±0.3</sup>
n=64	Qwen-max	89.5 <sup>±0.0</sup>	46.0 <sup>±0.0</sup>	41.5 <sup>±0.0</sup>	70.0 <sup>±0.0</sup>	76.0 <sup>±0.0</sup>	72.0 <sup>±0.0</sup>	64.5 <sup>±0.0</sup>	80.0 <sup>±0.0</sup>
	GPT-4	88.5 <sup>±1.0</sup>	45.4 <sup>±0.6</sup>	41.8 <sup>±0.3</sup>	70.5 <sup>±0.5</sup>	75.2 <sup>±0.8</sup>	71.5 <sup>±0.5</sup>	64.0 <sup>±0.5</sup>	79.2 <sup>±0.8</sup>
	Llama-7B	81.5 <sup>±8.0</sup>	35.4 <sup>±10.6</sup>	82.5 <sup>±41.0</sup>	60.6 <sup>±9.4</sup>	44.2 <sup>±31.8</sup>	42.9 <sup>±29.1</sup>	32.9 <sup>±31.6</sup>	40.9 <sup>±39.1</sup>
	Llama-13B	96.8 <sup>±7.3</sup>	40.0 <sup>±6.0</sup>	93.1 <sup>±51.6</sup>	82.5 <sup>±12.5</sup>	54.8 <sup>±21.2</sup>	52.8 <sup>±19.2</sup>	40.3 <sup>±24.2</sup>	49.8 <sup>±30.2</sup>
	Llama-70B	96.3 <sup>±6.8</sup>	58.0 <sup>±12.0</sup>	96.9 <sup>±55.4</sup>	92.0 <sup>±22.0</sup>	90.8 <sup>±14.8</sup>	88.9 <sup>±16.9</sup>	68.2 <sup>±3.7</sup>	82.2 <sup>±2.2</sup>

## G.2 NUMBER OF LLMs

We evaluate M2CL’s performance with varying number of LLMs, ranging from 4 to 64. The comparative results are shown in Tables 4 to 8 and Figs. 5 to 9.

**Summary of key findings.** The results show that M2CL significantly improves MAD performance, consistently surpassing existing methods by 20% – 50%, particularly in complex tasks like math and tool-using. This highlights its ability to tackle intricate reasoning tasks. M2CL also exhibits a more effective multi-agent scaling law, where performance consistently improves as the number of LLMs increases to 64, especially in agentic domains where a larger amount of LLMs enhances problem-solving accuracy and efficiency.

Table 4: Accuracy with varying number of LLMs on different datasets. The number of LLMs is 4 for all datasets. We exhibit the performance advantage with BoN and highlight the **best** result.

Model	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
Llama-7B	Single	45.3 <sub>±5.2</sub>	2.9 <sub>±4.0</sub>	16.8 <sub>±10.7</sub>	15.8 <sub>±6.0</sub>	22.5 <sub>±2.9</sub>	20.1 <sub>±4.8</sub>	15.3 <sub>±2.9</sub>	24.2 <sub>±2.0</sub>
	BoN	50.5 <sub>±0.0</sub>	6.9 <sub>±0.0</sub>	27.5 <sub>±0.0</sub>	21.8 <sub>±0.0</sub>	25.4 <sub>±0.0</sub>	24.9 <sub>±0.0</sub>	18.2 <sub>±0.0</sub>	26.2 <sub>±0.0</sub>
	Debate	47.3 <sub>±3.2</sub>	6.4 <sub>±0.5</sub>	26.2 <sub>±1.3</sub>	19.8 <sub>±2.0</sub>	25.3 <sub>±0.1</sub>	23.8 <sub>±1.1</sub>	16.5 <sub>±1.7</sub>	26.2 <sub>±0.0</sub>
	DyLAN	48.1 <sub>±2.4</sub>	6.6 <sub>±0.3</sub>	34.5 <sub>±7.0</sub>	22.2 <sub>±0.4</sub>	24.7 <sub>±0.7</sub>	21.8 <sub>±3.1</sub>	16.8 <sub>±1.4</sub>	25.5 <sub>±0.7</sub>
	GPTSwarm	48.0 <sub>±2.5</sub>	5.8 <sub>±1.1</sub>	26.0 <sub>±1.5</sub>	18.9 <sub>±2.9</sub>	25.8 <sub>±0.4</sub>	25.1 <sub>±0.2</sub>	18.8 <sub>±0.6</sub>	27.9 <sub>±1.7</sub>
	MacNet	50.8 <sub>±0.3</sub>	7.9 <sub>±1.0</sub>	30.7 <sub>±3.2</sub>	24.0 <sub>±2.2</sub>	27.8 <sub>±2.4</sub>	26.5 <sub>±1.6</sub>	18.8 <sub>±0.6</sub>	29.3 <sub>±3.1</sub>
	M2CL (ours)	59.1 <sub>±8.6</sub>	17.4 <sub>±10.5</sub>	44.7 <sub>±17.2</sub>	32.9 <sub>±11.1</sub>	35.1 <sub>±9.7</sub>	33.0 <sub>±8.1</sub>	25.5 <sub>±7.3</sub>	33.1 <sub>±6.9</sub>
Llama-14B	Single	54.8 <sub>±16.2</sub>	3.9 <sub>±4.9</sub>	25.0 <sub>±12.7</sub>	23.7 <sub>±8.3</sub>	28.5 <sub>±2.7</sub>	24.5 <sub>±7.2</sub>	18.6 <sub>±3.2</sub>	30.4 <sub>±2.8</sub>
	BoN	71.0 <sub>±0.0</sub>	8.8 <sub>±0.0</sub>	37.7 <sub>±0.0</sub>	32.0 <sub>±0.0</sub>	31.2 <sub>±0.0</sub>	31.7 <sub>±0.0</sub>	21.8 <sub>±0.0</sub>	33.2 <sub>±0.0</sub>
	Debate	64.5 <sub>±6.5</sub>	8.4 <sub>±0.4</sub>	36.3 <sub>±1.4</sub>	27.9 <sub>±4.1</sub>	30.8 <sub>±0.4</sub>	29.3 <sub>±2.4</sub>	19.8 <sub>±2.0</sub>	31.9 <sub>±1.3</sub>
	DyLAN	69.7 <sub>±1.3</sub>	10.3 <sub>±1.5</sub>	46.5 <sub>±8.8</sub>	32.5 <sub>±0.5</sub>	31.3 <sub>±0.1</sub>	28.7 <sub>±3.0</sub>	19.4 <sub>±2.4</sub>	30.4 <sub>±2.8</sub>
	MacNet	70.3 <sub>±0.7</sub>	9.2 <sub>±0.4</sub>	40.9 <sub>±3.2</sub>	35.2 <sub>±3.2</sub>	31.5 <sub>±0.3</sub>	31.8 <sub>±0.1</sub>	23.4 <sub>±1.6</sub>	35.3 <sub>±2.1</sub>
	MacNet	70.3 <sub>±0.7</sub>	9.2 <sub>±0.4</sub>	40.9 <sub>±3.2</sub>	35.2 <sub>±3.2</sub>	31.5 <sub>±0.3</sub>	31.8 <sub>±0.1</sub>	23.4 <sub>±1.6</sub>	35.3 <sub>±2.1</sub>
	M2CL (ours)	86.9 <sub>±15.9</sub>	23.4 <sub>±14.6</sub>	58.9 <sub>±21.2</sub>	47.5 <sub>±15.5</sub>	42.9 <sub>±11.7</sub>	39.9 <sub>±8.2</sub>	30.5 <sub>±8.7</sub>	40.9 <sub>±7.7</sub>
Llama-70B	Single	68.9 <sub>±14.1</sub>	31.6 <sub>±3.2</sub>	27.6 <sub>±32.0</sub>	35.5 <sub>±12.5</sub>	46.1 <sub>±4.8</sub>	40.2 <sub>±6.5</sub>	29.9 <sub>±2.8</sub>	48.4 <sub>±4.4</sub>
	BoN	83.0 <sub>±0.0</sub>	34.8 <sub>±0.0</sub>	59.6 <sub>±0.0</sub>	48.0 <sub>±0.0</sub>	50.9 <sub>±0.0</sub>	46.7 <sub>±0.0</sub>	32.7 <sub>±0.0</sub>	52.8 <sub>±0.0</sub>
	Debate	74.8 <sub>±8.2</sub>	29.8 <sub>±5.0</sub>	50.2 <sub>±9.4</sub>	43.2 <sub>±4.8</sub>	50.4 <sub>±0.5</sub>	47.8 <sub>±1.1</sub>	32.4 <sub>±0.3</sub>	52.9 <sub>±0.1</sub>
	DyLAN	82.1 <sub>±0.9</sub>	31.5 <sub>±3.3</sub>	53.8 <sub>±5.8</sub>	44.5 <sub>±3.5</sub>	49.3 <sub>±1.6</sub>	45.1 <sub>±1.6</sub>	31.5 <sub>±1.2</sub>	50.3 <sub>±2.5</sub>
	GPTSwarm	81.4 <sub>±1.6</sub>	31.5 <sub>±3.3</sub>	54.8 <sub>±4.8</sub>	43.7 <sub>±4.3</sub>	51.0 <sub>±0.1</sub>	45.2 <sub>±1.5</sub>	32.5 <sub>±0.2</sub>	51.5 <sub>±1.3</sub>
	MacNet	72.9 <sub>±10.1</sub>	32.7 <sub>±2.1</sub>	49.1 <sub>±10.5</sub>	40.8 <sub>±7.2</sub>	52.7 <sub>±1.8</sub>	55.9 <sub>±9.2</sub>	36.5 <sub>±3.8</sub>	57.0 <sub>±4.2</sub>
	M2CL (ours)	95.6 <sub>±12.6</sub>	40.3 <sub>±5.5</sub>	78.7 <sub>±19.1</sub>	70.6 <sub>±22.6</sub>	69.5 <sub>±18.6</sub>	65.5 <sub>±18.8</sub>	49.6 <sub>±16.9</sub>	66.5 <sub>±13.7</sub>
Qwen-7B	Single	61.2 <sub>±13.0</sub>	12.9 <sub>±12.0</sub>	20.2 <sub>±16.2</sub>	51.2 <sub>±11.3</sub>	24.3 <sub>±6.6</sub>	25.9 <sub>±6.6</sub>	15.3 <sub>±4.7</sub>	21.1 <sub>±4.1</sub>
	BoN	74.2 <sub>±0.0</sub>	24.9 <sub>±0.0</sub>	36.4 <sub>±0.0</sub>	62.5 <sub>±0.0</sub>	30.9 <sub>±0.0</sub>	32.5 <sub>±0.0</sub>	20.0 <sub>±0.0</sub>	25.2 <sub>±0.0</sub>
	Debate	69.3 <sub>±4.9</sub>	17.9 <sub>±7.0</sub>	28.7 <sub>±7.7</sub>	54.5 <sub>±8.0</sub>	29.5 <sub>±1.4</sub>	31.4 <sub>±1.1</sub>	20.6 <sub>±0.6</sub>	24.3 <sub>±0.9</sub>
	DyLAN	74.2 <sub>±0.0</sub>	23.3 <sub>±1.6</sub>	34.2 <sub>±2.2</sub>	62.5 <sub>±0.0</sub>	29.5 <sub>±1.4</sub>	28.3 <sub>±4.2</sub>	19.0 <sub>±1.0</sub>	23.0 <sub>±2.2</sub>
	GPTSwarm	74.5 <sub>±0.3</sub>	24.1 <sub>±0.8</sub>	37.2 <sub>±0.8</sub>	60.4 <sub>±2.1</sub>	29.4 <sub>±1.5</sub>	29.5 <sub>±3.0</sub>	20.0 <sub>±0.0</sub>	23.9 <sub>±1.3</sub>
	MacNet	68.0 <sub>±6.2</sub>	20.3 <sub>±4.6</sub>	25.0 <sub>±11.4</sub>	57.1 <sub>±5.4</sub>	29.8 <sub>±1.1</sub>	33.8 <sub>±1.3</sub>	20.5 <sub>±0.5</sub>	26.6 <sub>±1.4</sub>
	M2CL (ours)	88.7 <sub>±14.5</sub>	40.8 <sub>±15.9</sub>	58.4 <sub>±22.0</sub>	76.8 <sub>±14.3</sub>	36.7 <sub>±5.8</sub>	41.7 <sub>±9.2</sub>	30.2 <sub>±10.2</sub>	33.0 <sub>±7.8</sub>
Qwen-14B	Single	67.2 <sub>±12.5</sub>	21.6 <sub>±6.2</sub>	21.2 <sub>±11.6</sub>	56.7 <sub>±12.7</sub>	30.2 <sub>±6.0</sub>	31.7 <sub>±7.7</sub>	19.2 <sub>±4.6</sub>	26.4 <sub>±2.4</sub>
	BoN	79.7 <sub>±0.0</sub>	27.8 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	69.4 <sub>±0.0</sub>	36.2 <sub>±0.0</sub>	39.4 <sub>±0.0</sub>	23.8 <sub>±0.0</sub>	28.8 <sub>±0.0</sub>
	Debate	71.7 <sub>±8.0</sub>	24.7 <sub>±3.1</sub>	24.8 <sub>±8.0</sub>	62.7 <sub>±6.7</sub>	35.6 <sub>±0.6</sub>	38.5 <sub>±0.9</sub>	25.5 <sub>±1.7</sub>	29.9 <sub>±1.1</sub>
	DyLAN	77.5 <sub>±2.2</sub>	26.9 <sub>±0.9</sub>	31.9 <sub>±0.9</sub>	67.9 <sub>±1.5</sub>	33.6 <sub>±2.6</sub>	34.9 <sub>±4.5</sub>	22.7 <sub>±1.1</sub>	27.8 <sub>±1.0</sub>
	GPTSwarm	80.1 <sub>±0.4</sub>	26.9 <sub>±0.9</sub>	33.0 <sub>±0.2</sub>	68.1 <sub>±1.3</sub>	34.6 <sub>±1.6</sub>	34.8 <sub>±4.6</sub>	24.5 <sub>±0.7</sub>	28.3 <sub>±0.5</sub>
	MacNet	71.2 <sub>±8.5</sub>	24.6 <sub>±3.2</sub>	27.6 <sub>±5.2</sub>	62.9 <sub>±6.5</sub>	36.1 <sub>±0.1</sub>	41.4 <sub>±2.0</sub>	28.6 <sub>±4.8</sub>	29.4 <sub>±0.6</sub>
	M2CL (ours)	91.4 <sub>±11.7</sub>	43.6 <sub>±15.8</sub>	64.6 <sub>±31.8</sub>	86.5 <sub>±17.1</sub>	45.0 <sub>±8.8</sub>	51.3 <sub>±11.9</sub>	37.2 <sub>±13.4</sub>	39.7 <sub>±10.9</sub>
Qwen-72B	Single	72.5 <sub>±11.7</sub>	31.6 <sub>±9.4</sub>	34.9 <sub>±11.0</sub>	59.1 <sub>±13.1</sub>	49.3 <sub>±7.3</sub>	51.1 <sub>±8.7</sub>	31.4 <sub>±7.2</sub>	41.8 <sub>±8.0</sub>
	BoN	84.2 <sub>±0.0</sub>	51.0 <sub>±0.0</sub>	45.9 <sub>±0.0</sub>	72.2 <sub>±0.0</sub>	56.6 <sub>±0.0</sub>	59.8 <sub>±0.0</sub>	38.6 <sub>±0.0</sub>	49.8 <sub>±0.0</sub>
	Debate	76.6 <sub>±7.6</sub>	48.0 <sub>±3.0</sub>	39.6 <sub>±6.3</sub>	66.7 <sub>±5.5</sub>	58.3 <sub>±1.7</sub>	62.5 <sub>±2.7</sub>	40.7 <sub>±2.1</sub>	47.8 <sub>±2.0</sub>
	DyLAN	84.6 <sub>±0.4</sub>	61.5 <sub>±10.5</sub>	45.5 <sub>±0.4</sub>	71.3 <sub>±0.9</sub>	54.2 <sub>±2.4</sub>	57.8 <sub>±2.0</sub>	38.2 <sub>±0.4</sub>	45.9 <sub>±3.9</sub>
	GPTSwarm	84.9 <sub>±0.7</sub>	60.9 <sub>±9.9</sub>	46.6 <sub>±0.7</sub>	69.7 <sub>±2.5</sub>	55.2 <sub>±1.4</sub>	59.1 <sub>±0.7</sub>	39.0 <sub>±0.4</sub>	48.8 <sub>±1.0</sub>
	MacNet	80.3 <sub>±3.9</sub>	48.1 <sub>±2.9</sub>	41.2 <sub>±4.7</sub>	68.1 <sub>±4.1</sub>	57.2 <sub>±0.6</sub>	64.1 <sub>±4.3</sub>	42.0 <sub>±3.4</sub>	49.8 <sub>±0.0</sub>
	M2CL (ours)	93.5 <sub>±9.3</sub>	69.3 <sub>±18.3</sub>	73.1 <sub>±27.2</sub>	88.3 <sub>±16.1</sub>	73.9 <sub>±17.3</sub>	81.9 <sub>±22.1</sub>	60.3 <sub>±21.7</sub>	65.3 <sub>±15.5</sub>



Table 5: Accuracy with varying number of LLMs on different datasets. The number of LLMs is 8 for all datasets. We exhibit the performance advantage with BoN and highlight the best result.

Model	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
Llama-7B	Single	45.3 <sub>↓5.2</sub>	2.9 <sub>↓4.0</sub>	16.8 <sub>↓10.7</sub>	15.8 <sub>↓6.0</sub>	22.6 <sub>↓3.6</sub>	19.7 <sub>↓3.3</sub>	14.9 <sub>↓4.0</sub>	23.6 <sub>↓3.1</sub>
	BoN	50.5 <sub>↑0.0</sub>	6.9 <sub>↑0.0</sub>	27.5 <sub>↑0.0</sub>	21.8 <sub>↑0.0</sub>	26.2 <sub>↑0.0</sub>	25.0 <sub>↑0.0</sub>	18.9 <sub>↑0.0</sub>	26.7 <sub>↑0.0</sub>
	Debate	49.5 <sub>↓1.0</sub>	6.2 <sub>↓0.7</sub>	24.8 <sub>↓2.7</sub>	19.2 <sub>↓2.6</sub>	25.9 <sub>↓0.3</sub>	24.7 <sub>↓0.3</sub>	15.9 <sub>↓3.0</sub>	26.9 <sub>↑0.2</sub>
	DyLAN	54.2 <sub>↑3.7</sub>	9.7 <sub>↑2.8</sub>	35 <sub>↑7.5</sub>	26 <sub>↑4.2</sub>	25.0 <sub>↓1.2</sub>	22.0 <sub>↓3.0</sub>	17.2 <sub>↓1.7</sub>	25.4 <sub>↓1.3</sub>
	GPTSwarm	49.2 <sub>↓1.3</sub>	6.8 <sub>↓0.1</sub>	28.7 <sub>↑1.2</sub>	20.7 <sub>↓1.1</sub>	26.2 <sub>↑0.0</sub>	26.4 <sub>↑1.4</sub>	20.0 <sub>↑1.1</sub>	29.3 <sub>↑2.6</sub>
	MacNet	56.5 <sub>↑6.0</sub>	11.0 <sub>↑4.1</sub>	39.6 <sub>↑12.1</sub>	30.0 <sub>↑8.2</sub>	30.4 <sub>↑4.2</sub>	30.2 <sub>↑5.2</sub>	20.6 <sub>↑1.7</sub>	31.7 <sub>↑5.0</sub>
	M2CL (ours)	63.8 <sub>↑13.3</sub>	24.9 <sub>↑18.0</sub>	59.3 <sub>↑31.8</sub>	38.8 <sub>↑17.0</sub>	37.2 <sub>↑11.0</sub>	36.1 <sub>↑11.1</sub>	27.1 <sub>↑8.2</sub>	34.7 <sub>↑8.0</sub>
Llama-14B	Single	54.8 <sub>↓16.2</sub>	3.9 <sub>↓4.9</sub>	25.0 <sub>↓12.7</sub>	23.7 <sub>↓8.3</sub>	27.8 <sub>↓4.1</sub>	24.1 <sub>↓8.2</sub>	18.6 <sub>↓3.9</sub>	29.4 <sub>↓5.4</sub>
	BoN	71.0 <sub>↑0.0</sub>	8.8 <sub>↑0.0</sub>	37.7 <sub>↑0.0</sub>	32.0 <sub>↑0.0</sub>	31.9 <sub>↑0.0</sub>	32.3 <sub>↑0.0</sub>	22.5 <sub>↑0.0</sub>	34.8 <sub>↑0.0</sub>
	Debate	66.9 <sub>↓4.1</sub>	8.0 <sub>↓0.8</sub>	34.5 <sub>↓3.2</sub>	30.8 <sub>↓1.2</sub>	31.7 <sub>↓0.2</sub>	30.3 <sub>↓2.0</sub>	20.0 <sub>↓2.5</sub>	32.0 <sub>↓2.8</sub>
	DyLAN	82.1 <sub>↑11.1</sub>	12.4 <sub>↑3.6</sub>	46.7 <sub>↑9.0</sub>	37.5 <sub>↑5.5</sub>	31.7 <sub>↓0.2</sub>	29.6 <sub>↓2.7</sub>	19.4 <sub>↓3.1</sub>	31.3 <sub>↓3.5</sub>
	MacNet	81.7 <sub>↑10.7</sub>	14.2 <sub>↑5.4</sub>	53.9 <sub>↑16.2</sub>	43.3 <sub>↑11.3</sub>	33.4 <sub>↑1.5</sub>	35.6 <sub>↑3.3</sub>	24.9 <sub>↑2.4</sub>	37.0 <sub>↑2.2</sub>
	MacNet	81.7 <sub>↑10.7</sub>	14.2 <sub>↑5.4</sub>	53.9 <sub>↑16.2</sub>	43.3 <sub>↑11.3</sub>	33.4 <sub>↑1.5</sub>	35.6 <sub>↑3.3</sub>	24.9 <sub>↑2.4</sub>	37.0 <sub>↑2.2</sub>
	M2CL (ours)	92.7 <sub>↑21.7</sub>	23.0 <sub>↑14.2</sub>	70.7 <sub>↑33.0</sub>	55.4 <sub>↑23.4</sub>	46.1 <sub>↑14.2</sub>	44.7 <sub>↑12.4</sub>	33.6 <sub>↑11.1</sub>	43.8 <sub>↑9.0</sub>
Llama-70B	Single	68.9 <sub>↓14.1</sub>	31.6 <sub>↓3.2</sub>	27.6 <sub>↓32.0</sub>	35.5 <sub>↓12.5</sub>	44.1 <sub>↓8.3</sub>	39.1 <sub>↓9.7</sub>	29.1 <sub>↓3.8</sub>	47.1 <sub>↓5.7</sub>
	BoN	83.0 <sub>↑0.0</sub>	34.8 <sub>↑0.0</sub>	59.6 <sub>↑0.0</sub>	48.0 <sub>↑0.0</sub>	52.4 <sub>↑0.0</sub>	48.8 <sub>↑0.0</sub>	32.2 <sub>↑0.0</sub>	52.8 <sub>↑0.0</sub>
	Debate	82.2 <sub>↓0.8</sub>	34.3 <sub>↓0.5</sub>	58.1 <sub>↓1.5</sub>	47.1 <sub>↓0.9</sub>	51.7 <sub>↓0.7</sub>	50.1 <sub>↑1.3</sub>	32.2 <sub>↓0.7</sub>	53.3 <sub>↑0.5</sub>
	DyLAN	93.5 <sub>↑10.5</sub>	37.1 <sub>↑2.3</sub>	72.8 <sub>↑13.2</sub>	56.8 <sub>↑8.8</sub>	50.2 <sub>↓2.2</sub>	46.4 <sub>↓2.4</sub>	31.7 <sub>↓1.2</sub>	49.7 <sub>↓3.1</sub>
	GPTSwarm	93.6 <sub>↑10.6</sub>	36.6 <sub>↑1.8</sub>	71.2 <sub>↑11.6</sub>	57.1 <sub>↑9.1</sub>	51.4 <sub>↓1.0</sub>	46.5 <sub>↓2.3</sub>	32.2 <sub>↓0.7</sub>	52.2 <sub>↓0.6</sub>
	MacNet	82.0 <sub>↓1.0</sub>	36.9 <sub>↑2.1</sub>	62.2 <sub>↓2.6</sub>	48.8 <sub>↑0.8</sub>	57.3 <sub>↑4.9</sub>	64.0 <sub>↑15.2</sub>	40.7 <sub>↑7.8</sub>	62.6 <sub>↑9.8</sub>
	M2CL (ours)	95.4 <sub>↑12.4</sub>	43.5 <sub>↑8.7</sub>	87.6 <sub>↑28.0</sub>	81.0 <sub>↑33.0</sub>	78.9 <sub>↑26.5</sub>	76.2 <sub>↑27.4</sub>	58.7 <sub>↑25.8</sub>	73.6 <sub>↑20.8</sub>
Qwen-7B	Single	61.2 <sub>↓13.0</sub>	12.9 <sub>↓12.0</sub>	20.2 <sub>↓16.2</sub>	51.2 <sub>↓11.3</sub>	23.8 <sub>↓7.7</sub>	25.2 <sub>↓10.1</sub>	15.6 <sub>↓5.5</sub>	21.0 <sub>↓5.3</sub>
	BoN	74.2 <sub>↑0.0</sub>	24.9 <sub>↑0.0</sub>	36.4 <sub>↑0.0</sub>	62.5 <sub>↑0.0</sub>	31.5 <sub>↑0.0</sub>	35.3 <sub>↑0.0</sub>	21.1 <sub>↑0.0</sub>	26.3 <sub>↑0.0</sub>
	Debate	71.1 <sub>↓3.1</sub>	19.9 <sub>↓5.0</sub>	28.7 <sub>↓7.7</sub>	60.0 <sub>↓2.5</sub>	30.6 <sub>↓0.9</sub>	32.2 <sub>↓3.1</sub>	21.0 <sub>↓0.1</sub>	24.8 <sub>↓1.5</sub>
	DyLAN	74.3 <sub>↑0.1</sub>	26.7 <sub>↑1.8</sub>	35.4 <sub>↓1.0</sub>	63.4 <sub>↑0.9</sub>	29.8 <sub>↓1.7</sub>	29.4 <sub>↓5.9</sub>	18.4 <sub>↓2.7</sub>	23.4 <sub>↓2.9</sub>
	GPTSwarm	76.3 <sub>↑2.1</sub>	26.2 <sub>↑1.3</sub>	35.6 <sub>↓0.8</sub>	62.7 <sub>↑0.2</sub>	29.9 <sub>↓1.6</sub>	30.8 <sub>↓4.5</sub>	20.6 <sub>↓0.5</sub>	24.5 <sub>↓1.8</sub>
	MacNet	71.5 <sub>↓2.7</sub>	21.4 <sub>↓3.5</sub>	30.8 <sub>↓5.6</sub>	59.3 <sub>↓3.2</sub>	33.6 <sub>↑2.1</sub>	37.0 <sub>↑1.7</sub>	21.5 <sub>↑0.4</sub>	29.5 <sub>↑3.2</sub>
	M2CL (ours)	92.5 <sub>↑18.3</sub>	47.8 <sub>↑22.9</sub>	66.1 <sub>↑29.7</sub>	80.3 <sub>↑17.8</sub>	39.9 <sub>↑8.4</sub>	45.3 <sub>↑10.0</sub>	33.6 <sub>↑12.5</sub>	34.7 <sub>↑8.4</sub>
Qwen-14B	Single	67.2 <sub>↓12.5</sub>	21.6 <sub>↓6.2</sub>	21.2 <sub>↓11.6</sub>	56.7 <sub>↓12.7</sub>	30.1 <sub>↓7.4</sub>	31.3 <sub>↓10.3</sub>	18.7 <sub>↓6.9</sub>	25.7 <sub>↓2.8</sub>
	BoN	79.7 <sub>↑0.0</sub>	27.8 <sub>↑0.0</sub>	32.8 <sub>↑0.0</sub>	69.4 <sub>↑0.0</sub>	37.5 <sub>↑0.0</sub>	41.6 <sub>↑0.0</sub>	25.6 <sub>↑0.0</sub>	28.5 <sub>↑0.0</sub>
	Debate	77.2 <sub>↓2.5</sub>	27.4 <sub>↓0.4</sub>	30.3 <sub>↓2.5</sub>	66.9 <sub>↓2.5</sub>	36.8 <sub>↓0.7</sub>	39.4 <sub>↓2.2</sub>	26.4 <sub>↑0.8</sub>	30.0 <sub>↑1.5</sub>
	DyLAN	86.8 <sub>↑7.1</sub>	31.2 <sub>↑3.4</sub>	39.4 <sub>↑6.6</sub>	76.6 <sub>↑7.2</sub>	34.8 <sub>↓2.7</sub>	36.0 <sub>↓5.6</sub>	22.8 <sub>↓2.8</sub>	28.0 <sub>↓0.5</sub>
	GPTSwarm	87.0 <sub>↑7.3</sub>	31.3 <sub>↑3.5</sub>	38.7 <sub>↑5.9</sub>	75.9 <sub>↑6.5</sub>	34.9 <sub>↓2.6</sub>	35.9 <sub>↓5.7</sub>	25.0 <sub>↓0.6</sub>	28.4 <sub>↓0.1</sub>
	MacNet	78.9 <sub>↓0.8</sub>	28.7 <sub>↑0.9</sub>	31.9 <sub>↓0.9</sub>	70.4 <sub>↑1.0</sub>	39.2 <sub>↑1.7</sub>	46.0 <sub>↑4.4</sub>	32.2 <sub>↑6.6</sub>	32.7 <sub>↑4.2</sub>
	M2CL (ours)	93.7 <sub>↑14.0</sub>	51.7 <sub>↑23.9</sub>	66.2 <sub>↑33.4</sub>	91.1 <sub>↑21.7</sub>	48.2 <sub>↑10.7</sub>	56.1 <sub>↑14.5</sub>	42.0 <sub>↑16.4</sub>	43.0 <sub>↑14.5</sub>
Qwen-72B	Single	72.5 <sub>↓11.7</sub>	31.6 <sub>↓19.4</sub>	34.9 <sub>↓11.0</sub>	59.1 <sub>↓13.1</sub>	48.2 <sub>↓9.3</sub>	50.4 <sub>↓11.7</sub>	31.1 <sub>↓10.1</sub>	41.0 <sub>↓10.0</sub>
	BoN	84.2 <sub>↑0.0</sub>	51.0 <sub>↑0.0</sub>	45.9 <sub>↑0.0</sub>	72.2 <sub>↑0.0</sub>	57.5 <sub>↑0.0</sub>	62.1 <sub>↑0.0</sub>	41.2 <sub>↑0.0</sub>	51.0 <sub>↑0.0</sub>
	Debate	82.7 <sub>↓1.5</sub>	48.4 <sub>↓2.6</sub>	43.4 <sub>↓2.5</sub>	69.1 <sub>↓3.1</sub>	60.4 <sub>↑2.9</sub>	65.2 <sub>↑3.1</sub>	42.9 <sub>↑1.7</sub>	49.1 <sub>↓1.9</sub>
	DyLAN	91.5 <sub>↑7.3</sub>	63.1 <sub>↑12.1</sub>	51.6 <sub>↑5.7</sub>	80.4 <sub>↑8.2</sub>	55.1 <sub>↓2.4</sub>	58.0 <sub>↓4.1</sub>	40.4 <sub>↓0.8</sub>	45.5 <sub>↓5.5</sub>
	GPTSwarm	91.5 <sub>↑7.3</sub>	64.7 <sub>↑13.7</sub>	52.4 <sub>↑6.5</sub>	79.6 <sub>↑7.4</sub>	56.8 <sub>↓0.7</sub>	60.2 <sub>↓1.9</sub>	40.3 <sub>↓0.9</sub>	49.7 <sub>↓1.3</sub>
	MacNet	83.8 <sub>↓0.4</sub>	52.9 <sub>↑1.9</sub>	46.2 <sub>↑0.3</sub>	70.5 <sub>↓1.7</sub>	61.8 <sub>↑4.3</sub>	68.4 <sub>↑6.3</sub>	46.4 <sub>↑5.2</sub>	53.7 <sub>↑2.7</sub>
	M2CL (ours)	95.1 <sub>↑10.9</sub>	72.5 <sub>↑21.5</sub>	78.9 <sub>↑33.0</sub>	90.7 <sub>↑18.5</sub>	79.0 <sub>↑21.5</sub>	88.9 <sub>↑26.8</sub>	67.2 <sub>↑26.0</sub>	70.5 <sub>↑19.5</sub>

Table 6: Accuracy with varying number of LLMs on different datasets. The number of LLMs is 16 for all datasets. We exhibit the performance advantage with BoN and highlight the best result.

Model	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
Llama-7B	Single	45.3 <sub>±5.2</sub>	2.9 <sub>±4.0</sub>	16.8 <sub>±10.7</sub>	15.8 <sub>±6.0</sub>	22.4 <sub>±4.3</sub>	19.6 <sub>±6.1</sub>	14.7 <sub>±4.5</sub>	24.2 <sub>±2.1</sub>
	BoN	50.5 <sub>±0.0</sub>	6.9 <sub>±0.0</sub>	27.5 <sub>±0.0</sub>	21.8 <sub>±0.0</sub>	26.7 <sub>±0.0</sub>	25.7 <sub>±0.0</sub>	19.2 <sub>±0.0</sub>	26.3 <sub>±0.0</sub>
	Debate	51 <sub>±0.5</sub>	6.4 <sub>±0.5</sub>	26.1 <sub>±1.4</sub>	21.3 <sub>±0.5</sub>	25.7 <sub>±1.0</sub>	25.3 <sub>±0.4</sub>	16.6 <sub>±2.6</sub>	26.4 <sub>±0.1</sub>
	DyLAN	54.5 <sub>±4.0</sub>	9.9 <sub>±3.0</sub>	35.5 <sub>±8.0</sub>	26.2 <sub>±4.4</sub>	24.6 <sub>±2.1</sub>	22.7 <sub>±3.0</sub>	17.2 <sub>±2.0</sub>	25.8 <sub>±0.5</sub>
	GPTSwarm	51.6 <sub>±1.1</sub>	6.9 <sub>±0.0</sub>	29.1 <sub>±1.6</sub>	23.0 <sub>±1.2</sub>	27.2 <sub>±0.5</sub>	26.8 <sub>±1.1</sub>	20.8 <sub>±1.6</sub>	29.7 <sub>±3.4</sub>
	MacNet	59.3 <sub>±8.8</sub>	12.5 <sub>±5.6</sub>	45.0 <sub>±17.5</sub>	33.7 <sub>±11.9</sub>	32.8 <sub>±6.1</sub>	33.4 <sub>±7.7</sub>	22.6 <sub>±3.4</sub>	34.2 <sub>±7.9</sub>
	M2CL (ours)	71.5 <sub>±21.0</sub>	23.9 <sub>±17.0</sub>	69.9 <sub>±42.4</sub>	48.3 <sub>±26.5</sub>	39.8 <sub>±13.1</sub>	38.4 <sub>±12.7</sub>	28.9 <sub>±9.7</sub>	37.0 <sub>±0.7</sub>
Llama-14B	Single	54.8 <sub>±16.2</sub>	3.9 <sub>±4.9</sub>	25.0 <sub>±12.7</sub>	23.7 <sub>±8.3</sub>	27.4 <sub>±5.6</sub>	24.0 <sub>±9.5</sub>	18.3 <sub>±5.0</sub>	29.2 <sub>±4.6</sub>
	BoN	71.0 <sub>±0.0</sub>	8.8 <sub>±0.0</sub>	37.7 <sub>±0.0</sub>	32.0 <sub>±0.0</sub>	33.0 <sub>±0.0</sub>	33.5 <sub>±0.0</sub>	23.3 <sub>±0.0</sub>	33.8 <sub>±0.0</sub>
	Debate	72.2 <sub>±1.2</sub>	8.2 <sub>±0.6</sub>	37.0 <sub>±0.7</sub>	32.0 <sub>±0.0</sub>	32.0 <sub>±1.0</sub>	30.9 <sub>±2.6</sub>	19.6 <sub>±3.7</sub>	32.7 <sub>±1.1</sub>
	DyLAN	83.2 <sub>±12.2</sub>	12.4 <sub>±3.6</sub>	47.0 <sub>±9.3</sub>	38.1 <sub>±6.1</sub>	31.9 <sub>±1.1</sub>	30.2 <sub>±3.3</sub>	20.0 <sub>±3.3</sub>	31.2 <sub>±2.6</sub>
	MacNet	86.8 <sub>±15.8</sub>	16.0 <sub>±7.2</sub>	59.7 <sub>±22.0</sub>	49.1 <sub>±17.1</sub>	34.8 <sub>±1.8</sub>	37.8 <sub>±4.3</sub>	26.7 <sub>±3.4</sub>	39.9 <sub>±6.1</sub>
	MacNet	86.8 <sub>±15.8</sub>	16.0 <sub>±7.2</sub>	59.7 <sub>±22.0</sub>	49.1 <sub>±17.1</sub>	34.8 <sub>±1.8</sub>	37.8 <sub>±4.3</sub>	26.7 <sub>±3.4</sub>	39.9 <sub>±6.1</sub>
	M2CL (ours)	94.5 <sub>±23.5</sub>	32.7 <sub>±23.9</sub>	88.5 <sub>±50.8</sub>	66.1 <sub>±34.1</sub>	48.7 <sub>±15.7</sub>	47.1 <sub>±13.6</sub>	36.2 <sub>±12.9</sub>	45.5 <sub>±11.7</sub>
Llama-70B	Single	68.9 <sub>±14.1</sub>	31.6 <sub>±3.2</sub>	27.6 <sub>±32.0</sub>	35.5 <sub>±12.5</sub>	43.7 <sub>±10.2</sub>	38.3 <sub>±11.5</sub>	28.6 <sub>±4.6</sub>	46.8 <sub>±6.2</sub>
	BoN	83.0 <sub>±0.0</sub>	34.8 <sub>±0.0</sub>	59.6 <sub>±0.0</sub>	48.0 <sub>±0.0</sub>	53.9 <sub>±0.0</sub>	49.8 <sub>±0.0</sub>	33.2 <sub>±0.0</sub>	53.0 <sub>±0.0</sub>
	Debate	82.9 <sub>±0.1</sub>	34.6 <sub>±0.2</sub>	58.5 <sub>±1.1</sub>	46.8 <sub>±1.2</sub>	52.2 <sub>±1.7</sub>	51.2 <sub>±1.4</sub>	33.0 <sub>±0.2</sub>	52.8 <sub>±0.2</sub>
	DyLAN	93.2 <sub>±10.2</sub>	37.2 <sub>±2.4</sub>	73.2 <sub>±13.6</sub>	57.2 <sub>±9.2</sub>	50.5 <sub>±3.4</sub>	46.8 <sub>±3.0</sub>	31.1 <sub>±2.1</sub>	49.7 <sub>±3.3</sub>
	GPTSwarm	93.0 <sub>±10.0</sub>	37.4 <sub>±2.6</sub>	69.7 <sub>±10.1</sub>	55.5 <sub>±7.5</sub>	51.7 <sub>±2.2</sub>	47.5 <sub>±2.3</sub>	32.4 <sub>±0.8</sub>	51.6 <sub>±1.4</sub>
	MacNet	86.7 <sub>±3.7</sub>	38.9 <sub>±4.1</sub>	68.3 <sub>±8.7</sub>	49.8 <sub>±1.8</sub>	61.6 <sub>±7.7</sub>	70.8 <sub>±21.0</sub>	43.8 <sub>±10.6</sub>	67.1 <sub>±14.1</sub>
	M2CL (ours)	93.9 <sub>±10.9</sub>	51.5 <sub>±16.7</sub>	91.3 <sub>±31.7</sub>	97.2 <sub>±49.2</sub>	84.9 <sub>±31.0</sub>	81.7 <sub>±31.9</sub>	61.5 <sub>±28.3</sub>	78.1 <sub>±25.1</sub>
Qwen-7B	Single	61.2 <sub>±13.0</sub>	12.9 <sub>±12.0</sub>	20.2 <sub>±16.2</sub>	51.2 <sub>±11.3</sub>	23.4 <sub>±9.5</sub>	24.9 <sub>±10.5</sub>	15.6 <sub>±6.4</sub>	20.6 <sub>±6.5</sub>
	BoN	74.2 <sub>±0.0</sub>	24.9 <sub>±0.0</sub>	36.4 <sub>±0.0</sub>	62.5 <sub>±0.0</sub>	32.9 <sub>±0.0</sub>	35.4 <sub>±0.0</sub>	22.0 <sub>±0.0</sub>	27.1 <sub>±0.0</sub>
	Debate	73.4 <sub>±0.8</sub>	24.1 <sub>±0.8</sub>	35.6 <sub>±0.8</sub>	62.1 <sub>±0.4</sub>	31.7 <sub>±1.2</sub>	32.6 <sub>±2.8</sub>	22.1 <sub>±0.1</sub>	24.2 <sub>±2.9</sub>
	DyLAN	77.4 <sub>±3.2</sub>	26.0 <sub>±1.1</sub>	38.9 <sub>±2.5</sub>	64.8 <sub>±2.3</sub>	30.4 <sub>±2.5</sub>	28.7 <sub>±6.7</sub>	19.2 <sub>±2.8</sub>	22.7 <sub>±4.4</sub>
	GPTSwarm	78.0 <sub>±3.8</sub>	27.6 <sub>±2.7</sub>	41.1 <sub>±4.7</sub>	66.2 <sub>±3.7</sub>	30.1 <sub>±2.8</sub>	30.4 <sub>±5.0</sub>	22.1 <sub>±0.1</sub>	24.4 <sub>±2.7</sub>
	MacNet	76.8 <sub>±2.6</sub>	27.8 <sub>±2.9</sub>	40.3 <sub>±3.9</sub>	69.4 <sub>±6.9</sub>	35.4 <sub>±2.5</sub>	39.7 <sub>±4.3</sub>	24.6 <sub>±2.6</sub>	30.7 <sub>±3.6</sub>
	M2CL (ours)	94.6 <sub>±20.4</sub>	47.4 <sub>±22.5</sub>	72.8 <sub>±36.4</sub>	84.9 <sub>±22.4</sub>	41.8 <sub>±8.9</sub>	47.6 <sub>±12.2</sub>	36.5 <sub>±14.5</sub>	37.4 <sub>±10.3</sub>
Qwen-14B	Single	67.2 <sub>±12.5</sub>	21.6 <sub>±6.2</sub>	21.2 <sub>±11.6</sub>	56.7 <sub>±12.7</sub>	28.3 <sub>±9.8</sub>	30.9 <sub>±11.4</sub>	18.5 <sub>±7.6</sub>	25.5 <sub>±3.8</sub>
	BoN	79.7 <sub>±0.0</sub>	27.8 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	69.4 <sub>±0.0</sub>	38.1 <sub>±0.0</sub>	42.3 <sub>±0.0</sub>	26.1 <sub>±0.0</sub>	29.3 <sub>±0.0</sub>
	Debate	78.7 <sub>±1.0</sub>	26.7 <sub>±1.1</sub>	33.4 <sub>±0.6</sub>	70.1 <sub>±0.7</sub>	37.7 <sub>±0.4</sub>	40.5 <sub>±1.8</sub>	26.7 <sub>±0.6</sub>	30.4 <sub>±1.1</sub>
	DyLAN	88.0 <sub>±8.3</sub>	31.9 <sub>±4.1</sub>	41.0 <sub>±8.2</sub>	78.2 <sub>±8.8</sub>	34.9 <sub>±3.2</sub>	35.7 <sub>±6.6</sub>	23.9 <sub>±2.2</sub>	28.3 <sub>±1.0</sub>
	GPTSwarm	88.0 <sub>±8.3</sub>	31.8 <sub>±4.0</sub>	40.4 <sub>±7.6</sub>	78.4 <sub>±9.0</sub>	35.5 <sub>±2.6</sub>	36.8 <sub>±5.5</sub>	25.1 <sub>±1.0</sub>	28.7 <sub>±0.6</sub>
	MacNet	83.1 <sub>±3.4</sub>	30.8 <sub>±3.0</sub>	35.1 <sub>±2.3</sub>	74.8 <sub>±5.4</sub>	42.1 <sub>±4.0</sub>	49.3 <sub>±7.0</sub>	35.5 <sub>±9.4</sub>	33.8 <sub>±4.5</sub>
	M2CL (ours)	95.9 <sub>±16.2</sub>	52.7 <sub>±24.9</sub>	69.0 <sub>±36.2</sub>	93.3 <sub>±23.9</sub>	51.2 <sub>±13.1</sub>	59.4 <sub>±17.1</sub>	43.9 <sub>±17.8</sub>	46.0 <sub>±16.7</sub>
Qwen-72B	Single	72.5 <sub>±11.7</sub>	31.6 <sub>±19.4</sub>	34.9 <sub>±11.0</sub>	59.1 <sub>±13.1</sub>	47.7 <sub>±10.8</sub>	50.1 <sub>±12.8</sub>	31.2 <sub>±12.1</sub>	40.8 <sub>±11.5</sub>
	BoN	84.2 <sub>±0.0</sub>	51.0 <sub>±0.0</sub>	45.9 <sub>±0.0</sub>	72.2 <sub>±0.0</sub>	58.5 <sub>±0.0</sub>	62.9 <sub>±0.0</sub>	43.3 <sub>±0.0</sub>	52.3 <sub>±0.0</sub>
	Debate	83.9 <sub>±0.3</sub>	56.1 <sub>±5.1</sub>	44.1 <sub>±1.8</sub>	71.6 <sub>±0.6</sub>	60.8 <sub>±2.3</sub>	66.7 <sub>±3.8</sub>	44.1 <sub>±0.8</sub>	49.3 <sub>±3.0</sub>
	DyLAN	92.7 <sub>±8.5</sub>	65.8 <sub>±14.8</sub>	53.6 <sub>±7.7</sub>	81.4 <sub>±9.2</sub>	56.1 <sub>±2.4</sub>	59.1 <sub>±3.8</sub>	41.1 <sub>±2.2</sub>	46.3 <sub>±6.0</sub>
	GPTSwarm	92.4 <sub>±8.2</sub>	66.0 <sub>±15.0</sub>	53.3 <sub>±7.4</sub>	81.1 <sub>±8.9</sub>	57.3 <sub>±1.2</sub>	59.8 <sub>±3.1</sub>	41.3 <sub>±2.0</sub>	50.6 <sub>±1.7</sub>
	MacNet	88.1 <sub>±3.9</sub>	60.6 <sub>±9.6</sub>	49.2 <sub>±3.3</sub>	72.6 <sub>±0.4</sub>	65.8 <sub>±7.3</sub>	74.7 <sub>±11.8</sub>	50.3 <sub>±7.0</sub>	58.1 <sub>±5.8</sub>
	M2CL (ours)	96.6 <sub>±12.4</sub>	74.3 <sub>±23.3</sub>	80.0 <sub>±34.1</sub>	94.4 <sub>±22.2</sub>	82.9 <sub>±24.4</sub>	94.6 <sub>±31.7</sub>	71.9 <sub>±28.6</sub>	74.3 <sub>±22.0</sub>

Table 7: Accuracy with varying number of LLMs on different datasets. The number of LLMs is 32 for all datasets. We exhibit the performance advantage with BoN and highlight the best result.

Model	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
Llama-7B	Single	45.3 <sub>±5.2</sub>	2.9 <sub>±4.0</sub>	16.8 <sub>±10.7</sub>	15.8 <sub>±6.0</sub>	22.4 <sub>±4.4</sub>	18.8 <sub>±8.1</sub>	14.7 <sub>±5.1</sub>	23.2 <sub>±4.0</sub>
	BoN	50.5 <sub>±0.0</sub>	6.9 <sub>±0.0</sub>	27.5 <sub>±0.0</sub>	21.8 <sub>±0.0</sub>	26.8 <sub>±0.0</sub>	26.9 <sub>±0.0</sub>	19.8 <sub>±0.0</sub>	27.2 <sub>±0.0</sub>
	Debate	51.8 <sub>±1.3</sub>	7.7 <sub>±0.8</sub>	29.5 <sub>±2.0</sub>	23.5 <sub>±1.7</sub>	26.4 <sub>±0.4</sub>	25.6 <sub>±1.3</sub>	15.8 <sub>±4.0</sub>	26.2 <sub>±1.0</sub>
	DyLAN	54.7 <sub>±4.2</sub>	10.1 <sub>±3.2</sub>	36.2 <sub>±8.7</sub>	26.7 <sub>±4.9</sub>	24.9 <sub>±1.9</sub>	23.1 <sub>±3.8</sub>	17.6 <sub>±2.2</sub>	26.0 <sub>±1.2</sub>
	GPTSwarm	52.4 <sub>±1.9</sub>	7.9 <sub>±1.0</sub>	30.8 <sub>±3.3</sub>	23.7 <sub>±1.9</sub>	27.3 <sub>±0.5</sub>	27.9 <sub>±1.0</sub>	21.8 <sub>±2.0</sub>	30.4 <sub>±3.2</sub>
	MacNet	62.7 <sub>±12.2</sub>	14.0 <sub>±7.1</sub>	50.3 <sub>±22.8</sub>	37.2 <sub>±15.4</sub>	35.7 <sub>±8.9</sub>	37.5 <sub>±10.6</sub>	23.8 <sub>±4.0</sub>	37.5 <sub>±10.3</sub>
	M2CL (ours)	79.1 <sub>±28.6</sub>	28.8 <sub>±21.9</sub>	84.5 <sub>±57.0</sub>	56.5 <sub>±34.7</sub>	42.5 <sub>±15.7</sub>	40.7 <sub>±13.8</sub>	31.4 <sub>±11.6</sub>	38.9 <sub>±11.7</sub>
Llama-14B	Single	54.8 <sub>±16.2</sub>	3.9 <sub>±4.9</sub>	25.0 <sub>±12.7</sub>	23.7 <sub>±8.3</sub>	27.0 <sub>±5.8</sub>	23.5 <sub>±11.3</sub>	18.1 <sub>±6.1</sub>	28.9 <sub>±5.4</sub>
	BoN	71.0 <sub>±0.0</sub>	8.8 <sub>±0.0</sub>	37.7 <sub>±0.0</sub>	32.0 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	34.8 <sub>±0.0</sub>	24.2 <sub>±0.0</sub>	34.3 <sub>±0.0</sub>
	Debate	73.1 <sub>±2.1</sub>	9.5 <sub>±0.7</sub>	40.7 <sub>±3.0</sub>	34.3 <sub>±2.3</sub>	31.9 <sub>±0.9</sub>	31.7 <sub>±3.1</sub>	19.9 <sub>±4.3</sub>	33.0 <sub>±1.3</sub>
	DyLAN	83.6 <sub>±12.6</sub>	12.7 <sub>±3.9</sub>	47.7 <sub>±10.0</sub>	38.7 <sub>±6.7</sub>	32.1 <sub>±0.7</sub>	30.8 <sub>±4.0</sub>	19.8 <sub>±4.4</sub>	31.0 <sub>±3.3</sub>
	MacNet	90.7 <sub>±19.7</sub>	17.6 <sub>±8.8</sub>	67.2 <sub>±29.5</sub>	53.9 <sub>±21.9</sub>	37.1 <sub>±4.3</sub>	41.7 <sub>±6.9</sub>	29.1 <sub>±4.9</sub>	42.3 <sub>±8.0</sub>
	MacNet	90.7 <sub>±19.7</sub>	17.6 <sub>±8.8</sub>	67.2 <sub>±29.5</sub>	53.9 <sub>±21.9</sub>	37.1 <sub>±4.3</sub>	41.7 <sub>±6.9</sub>	29.1 <sub>±4.9</sub>	42.3 <sub>±8.0</sub>
	M2CL (ours)	95.8 <sub>±24.8</sub>	41.4 <sub>±32.6</sub>	94.7 <sub>±57.0</sub>	75.6 <sub>±43.6</sub>	52.2 <sub>±19.4</sub>	50.7 <sub>±15.9</sub>	38.5 <sub>±14.3</sub>	48.0 <sub>±13.7</sub>
Llama-70B	Single	68.9 <sub>±14.1</sub>	31.6 <sub>±3.2</sub>	27.6 <sub>±32.0</sub>	35.5 <sub>±12.5</sub>	43.1 <sub>±10.0</sub>	38.6 <sub>±11.5</sub>	28.1 <sub>±4.7</sub>	46.0 <sub>±7.2</sub>
	BoN	83.0 <sub>±0.0</sub>	34.8 <sub>±0.0</sub>	59.6 <sub>±0.0</sub>	48.0 <sub>±0.0</sub>	53.1 <sub>±0.0</sub>	50.1 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	53.2 <sub>±0.0</sub>
	Debate	85.7 <sub>±2.7</sub>	35.3 <sub>±0.5</sub>	63.4 <sub>±3.8</sub>	51.2 <sub>±3.2</sub>	52.6 <sub>±0.5</sub>	52.1 <sub>±2.0</sub>	32.9 <sub>±0.1</sub>	53.7 <sub>±0.5</sub>
	DyLAN	94.5 <sub>±11.5</sub>	37.3 <sub>±2.5</sub>	74.6 <sub>±15.0</sub>	57.8 <sub>±9.8</sub>	50.1 <sub>±3.0</sub>	47.2 <sub>±2.9</sub>	31.4 <sub>±1.4</sub>	49.7 <sub>±3.5</sub>
	GPTSwarm	96.4 <sub>±13.4</sub>	37.8 <sub>±3.0</sub>	73.4 <sub>±13.8</sub>	56.4 <sub>±8.4</sub>	52.2 <sub>±0.9</sub>	48.4 <sub>±1.7</sub>	32.5 <sub>±0.3</sub>	51.6 <sub>±1.6</sub>
	MacNet	90.0 <sub>±7.0</sub>	41.2 <sub>±6.4</sub>	73.9 <sub>±14.3</sub>	54.1 <sub>±6.1</sub>	64.5 <sub>±11.4</sub>	76.6 <sub>±26.5</sub>	46.5 <sub>±13.7</sub>	71.6 <sub>±18.4</sub>
	M2CL (ours)	97.0 <sub>±14.0</sub>	54.5 <sub>±19.7</sub>	95.1 <sub>±35.5</sub>	93.7 <sub>±45.7</sub>	88.5 <sub>±35.4</sub>	86.0 <sub>±35.9</sub>	65.7 <sub>±32.9</sub>	79.8 <sub>±26.6</sub>
Qwen-7B	Single	61.2 <sub>±13.0</sub>	12.9 <sub>±12.0</sub>	20.2 <sub>±16.2</sub>	51.2 <sub>±11.3</sub>	24.2 <sub>±8.6</sub>	24.5 <sub>±13.3</sub>	14.6 <sub>±7.7</sub>	20.4 <sub>±7.3</sub>
	BoN	74.2 <sub>±0.0</sub>	24.9 <sub>±0.0</sub>	36.4 <sub>±0.0</sub>	62.5 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	37.8 <sub>±0.0</sub>	22.3 <sub>±0.0</sub>	27.7 <sub>±0.0</sub>
	Debate	76.8 <sub>±2.6</sub>	26.8 <sub>±1.9</sub>	40.0 <sub>±3.6</sub>	64.3 <sub>±1.8</sub>	31.7 <sub>±1.1</sub>	33.4 <sub>±4.4</sub>	23.0 <sub>±0.7</sub>	25.3 <sub>±2.4</sub>
	DyLAN	78.2 <sub>±4.0</sub>	29.5 <sub>±4.6</sub>	40.3 <sub>±3.9</sub>	66.5 <sub>±4.0</sub>	30.7 <sub>±2.1</sub>	29.2 <sub>±8.6</sub>	19.0 <sub>±3.3</sub>	22.6 <sub>±5.1</sub>
	GPTSwarm	79.6 <sub>±5.4</sub>	29.9 <sub>±5.0</sub>	41.8 <sub>±5.4</sub>	67.8 <sub>±5.3</sub>	31.2 <sub>±1.6</sub>	31.3 <sub>±6.5</sub>	22.6 <sub>±0.3</sub>	25.0 <sub>±2.7</sub>
	MacNet	80.6 <sub>±6.4</sub>	31.5 <sub>±6.6</sub>	45.8 <sub>±9.4</sub>	73.7 <sub>±11.2</sub>	38.6 <sub>±5.8</sub>	43.1 <sub>±5.3</sub>	26.6 <sub>±4.3</sub>	34.3 <sub>±6.6</sub>
	M2CL (ours)	96.0 <sub>±21.8</sub>	47.1 <sub>±22.2</sub>	74.9 <sub>±38.5</sub>	87.3 <sub>±24.8</sub>	44.4 <sub>±11.6</sub>	50.4 <sub>±12.6</sub>	39.7 <sub>±17.4</sub>	40.2 <sub>±12.5</sub>
Qwen-14B	Single	67.2 <sub>±12.5</sub>	21.6 <sub>±6.2</sub>	21.2 <sub>±11.6</sub>	56.7 <sub>±12.7</sub>	29.4 <sub>±8.0</sub>	29.5 <sub>±15.6</sub>	18.9 <sub>±8.2</sub>	24.5 <sub>±4.5</sub>
	BoN	79.7 <sub>±0.0</sub>	27.8 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	69.4 <sub>±0.0</sub>	37.4 <sub>±0.0</sub>	45.1 <sub>±0.0</sub>	27.1 <sub>±0.0</sub>	29.0 <sub>±0.0</sub>
	Debate	82.7 <sub>±3.0</sub>	29.1 <sub>±1.3</sub>	35.5 <sub>±2.7</sub>	71.8 <sub>±2.4</sub>	39.0 <sub>±1.6</sub>	41.7 <sub>±3.4</sub>	28.7 <sub>±1.6</sub>	30.8 <sub>±1.8</sub>
	DyLAN	89.3 <sub>±9.6</sub>	32.6 <sub>±4.8</sub>	42.2 <sub>±9.4</sub>	78.9 <sub>±9.5</sub>	35.4 <sub>±2.0</sub>	36.2 <sub>±8.9</sub>	23.2 <sub>±3.9</sub>	29.3 <sub>±0.3</sub>
	GPTSwarm	89.0 <sub>±9.3</sub>	32.6 <sub>±4.8</sub>	42.2 <sub>±9.4</sub>	78.9 <sub>±9.5</sub>	36.4 <sub>±1.0</sub>	37.5 <sub>±7.6</sub>	26.6 <sub>±0.5</sub>	28.9 <sub>±0.1</sub>
	MacNet	87.8 <sub>±8.1</sub>	33.7 <sub>±5.9</sub>	40.1 <sub>±7.3</sub>	83.0 <sub>±13.6</sub>	44.8 <sub>±7.4</sub>	54.2 <sub>±9.1</sub>	39.8 <sub>±12.7</sub>	36.2 <sub>±7.2</sub>
	M2CL (ours)	96.1 <sub>±16.4</sub>	58.0 <sub>±30.2</sub>	70.5 <sub>±37.7</sub>	94.9 <sub>±25.5</sub>	54.9 <sub>±17.5</sub>	62.1 <sub>±17.0</sub>	49.0 <sub>±21.9</sub>	49.6 <sub>±20.6</sub>
Qwen-72B	Single	72.5 <sub>±11.7</sub>	31.6 <sub>±19.4</sub>	34.9 <sub>±11.0</sub>	59.1 <sub>±13.1</sub>	46.7 <sub>±12.6</sub>	49.0 <sub>±14.2</sub>	30.4 <sub>±13.5</sub>	40.1 <sub>±13.2</sub>
	BoN	84.2 <sub>±0.0</sub>	51.0 <sub>±0.0</sub>	45.9 <sub>±0.0</sub>	72.2 <sub>±0.0</sub>	59.3 <sub>±0.0</sub>	63.2 <sub>±0.0</sub>	43.9 <sub>±0.0</sub>	53.3 <sub>±0.0</sub>
	Debate	86.5 <sub>±2.3</sub>	62.9 <sub>±11.9</sub>	48.3 <sub>±2.4</sub>	74.5 <sub>±2.3</sub>	62.8 <sub>±3.5</sub>	67.8 <sub>±4.6</sub>	46.6 <sub>±2.7</sub>	50.2 <sub>±3.1</sub>
	DyLAN	92.8 <sub>±8.6</sub>	67.7 <sub>±16.7</sub>	54.8 <sub>±8.9</sub>	82.7 <sub>±10.5</sub>	56.0 <sub>±3.3</sub>	59.1 <sub>±4.1</sub>	42.9 <sub>±1.0</sub>	46.7 <sub>±6.6</sub>
	GPTSwarm	93.1 <sub>±8.9</sub>	68.1 <sub>±17.1</sub>	54.7 <sub>±8.8</sub>	82.2 <sub>±10.0</sub>	58.3 <sub>±1.0</sub>	62.6 <sub>±0.6</sub>	42.7 <sub>±1.2</sub>	51.4 <sub>±1.9</sub>
	MacNet	92.5 <sub>±8.3</sub>	72.1 <sub>±21.1</sub>	55.2 <sub>±9.3</sub>	79.9 <sub>±7.7</sub>	70.5 <sub>±11.2</sub>	79.5 <sub>±16.3</sub>	54.7 <sub>±10.8</sub>	62.3 <sub>±9.0</sub>
	M2CL (ours)	97.5 <sub>±13.3</sub>	74.7 <sub>±23.7</sub>	82.0 <sub>±36.1</sub>	95.9 <sub>±23.7</sub>	89.3 <sub>±30.0</sub>	95 <sub>±31.8</sub>	79.7 <sub>±35.8</sub>	80.0 <sub>±26.7</sub>

Table 8: Accuracy with varying number of LLMs on different datasets. The number of LLMs is 64 for all datasets. We exhibit the performance advantage with BoN and highlight the best result.

Model	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
Llama-7B	Single	45.3 <sub>±5.2</sub>	2.9 <sub>±4.0</sub>	16.8 <sub>±10.7</sub>	15.8 <sub>±6.0</sub>	22.0 <sub>±5.2</sub>	18.8 <sub>±7.9</sub>	15.0 <sub>±4.8</sub>	22.5 <sub>±4.4</sub>
	BoN	50.5 <sub>±0.0</sub>	6.9 <sub>±0.0</sub>	27.5 <sub>±0.0</sub>	21.8 <sub>±0.0</sub>	27.2 <sub>±0.0</sub>	26.7 <sub>±0.0</sub>	19.8 <sub>±0.0</sub>	26.9 <sub>±0.0</sub>
	Debate	53.4 <sub>±2.9</sub>	9.2 <sub>±2.3</sub>	33.5 <sub>±6.0</sub>	25 <sub>±3.2</sub>	26.9 <sub>±0.3</sub>	26.3 <sub>±0.4</sub>	16.5 <sub>±3.3</sub>	26.9 <sub>±0.0</sub>
	DyLAN	55.1 <sub>±4.6</sub>	10.4 <sub>±3.5</sub>	36.9 <sub>±9.4</sub>	27 <sub>±5.2</sub>	24.6 <sub>±2.6</sub>	23.3 <sub>±3.4</sub>	17.9 <sub>±1.9</sub>	25.0 <sub>±1.9</sub>
	GPTSwarm	53.5 <sub>±3.0</sub>	9 <sub>±2.1</sub>	34 <sub>±6.5</sub>	25.3 <sub>±3.5</sub>	27.0 <sub>±0.2</sub>	28.8 <sub>±2.1</sub>	22.1 <sub>±2.3</sub>	31.0 <sub>±4.1</sub>
	MacNet	66.4 <sub>±15.9</sub>	15.3 <sub>±8.4</sub>	54.4 <sub>±26.9</sub>	40.0 <sub>±18.2</sub>	38.8 <sub>±11.6</sub>	39.6 <sub>±12.9</sub>	27.0 <sub>±7.2</sub>	38.4 <sub>±11.5</sub>
	M2CL (ours)	81.5 <sub>±31.0</sub>	35.4 <sub>±28.5</sub>	82.5 <sub>±55.0</sub>	60.6 <sub>±38.8</sub>	44.2 <sub>±17.0</sub>	42.9 <sub>±16.2</sub>	32.9 <sub>±13.1</sub>	40.9 <sub>±14.0</sub>
Llama-14B	Single	54.8 <sub>±16.2</sub>	3.9 <sub>±4.9</sub>	25.0 <sub>±12.7</sub>	23.7 <sub>±8.3</sub>	26.9 <sub>±5.7</sub>	23.6 <sub>±12.1</sub>	17.8 <sub>±6.7</sub>	28.6 <sub>±5.6</sub>
	BoN	71.0 <sub>±0.0</sub>	8.8 <sub>±0.0</sub>	37.7 <sub>±0.0</sub>	32.0 <sub>±0.0</sub>	32.6 <sub>±0.0</sub>	35.7 <sub>±0.0</sub>	24.5 <sub>±0.0</sub>	34.2 <sub>±0.0</sub>
	Debate	79.6 <sub>±8.6</sub>	11.1 <sub>±2.3</sub>	45.0 <sub>±7.3</sub>	36.5 <sub>±4.5</sub>	32.6 <sub>±0.0</sub>	32.0 <sub>±3.7</sub>	19.9 <sub>±4.6</sub>	32.6 <sub>±1.6</sub>
	DyLAN	85.0 <sub>±14.0</sub>	13.0 <sub>±4.2</sub>	48.8 <sub>±11.1</sub>	39.3 <sub>±7.3</sub>	31.9 <sub>±0.7</sub>	31.0 <sub>±4.7</sub>	19.9 <sub>±4.6</sub>	30.5 <sub>±3.7</sub>
	MacNet	90.8 <sub>±19.8</sub>	18.1 <sub>±9.3</sub>	67.5 <sub>±29.8</sub>	55.2 <sub>±23.2</sub>	39.1 <sub>±6.5</sub>	44.4 <sub>±8.7</sub>	31.1 <sub>±6.6</sub>	44.0 <sub>±9.8</sub>
	MacNet	90.8 <sub>±19.8</sub>	18.1 <sub>±9.3</sub>	67.5 <sub>±29.8</sub>	55.2 <sub>±23.2</sub>	39.1 <sub>±6.5</sub>	44.4 <sub>±8.7</sub>	31.1 <sub>±6.6</sub>	44.0 <sub>±9.8</sub>
	M2CL (ours)	96.8 <sub>±25.8</sub>	40.0 <sub>±31.2</sub>	93.1 <sub>±55.4</sub>	82.5 <sub>±50.5</sub>	54.8 <sub>±22.2</sub>	52.8 <sub>±17.1</sub>	40.3 <sub>±15.8</sub>	49.8 <sub>±15.6</sub>
Llama-70B	Single	68.9 <sub>±14.1</sub>	31.6 <sub>±3.2</sub>	27.6 <sub>±32.0</sub>	35.5 <sub>±12.5</sub>	43.0 <sub>±11.8</sub>	37.3 <sub>±13.7</sub>	28.2 <sub>±5.7</sub>	45.9 <sub>±6.8</sub>
	BoN	83.0 <sub>±0.0</sub>	34.8 <sub>±0.0</sub>	59.6 <sub>±0.0</sub>	48.0 <sub>±0.0</sub>	54.8 <sub>±0.0</sub>	51.0 <sub>±0.0</sub>	33.9 <sub>±0.0</sub>	52.7 <sub>±0.0</sub>
	Debate	90.0 <sub>±7.0</sub>	36.6 <sub>±1.8</sub>	69.0 <sub>±9.4</sub>	54.0 <sub>±6.0</sub>	53.6 <sub>±1.2</sub>	53.5 <sub>±2.5</sub>	32.7 <sub>±1.2</sub>	53.2 <sub>±0.5</sub>
	DyLAN	95.4 <sub>±12.4</sub>	37.6 <sub>±2.8</sub>	75.4 <sub>±15.8</sub>	58.9 <sub>±10.9</sub>	50.4 <sub>±4.4</sub>	47.6 <sub>±3.4</sub>	30.8 <sub>±3.1</sub>	50.4 <sub>±2.3</sub>
	GPTSwarm	95.7 <sub>±12.7</sub>	37.9 <sub>±3.1</sub>	78.9 <sub>±9.3</sub>	58.1 <sub>±10.1</sub>	52.4 <sub>±2.4</sub>	49.3 <sub>±1.7</sub>	32.3 <sub>±1.6</sub>	51.3 <sub>±1.4</sub>
	MacNet	95.3 <sub>±12.3</sub>	44.2 <sub>±9.4</sub>	84.2 <sub>±24.6</sub>	57.9 <sub>±9.9</sub>	67.9 <sub>±13.1</sub>	81.6 <sub>±30.6</sub>	49.0 <sub>±15.1</sub>	75.4 <sub>±22.7</sub>
	M2CL (ours)	96.3 <sub>±13.3</sub>	58.0 <sub>±23.2</sub>	96.9 <sub>±37.3</sub>	92.0 <sub>±44.0</sub>	90.8 <sub>±36.0</sub>	88.9 <sub>±37.9</sub>	68.2 <sub>±34.3</sub>	82.2 <sub>±29.5</sub>
Qwen-7B	Single	61.2 <sub>±13.0</sub>	12.9 <sub>±12.0</sub>	20.2 <sub>±16.2</sub>	51.2 <sub>±11.3</sub>	23.0 <sub>±11.5</sub>	24.4 <sub>±14.2</sub>	14.8 <sub>±8.2</sub>	20.0 <sub>±7.8</sub>
	BoN	74.2 <sub>±0.0</sub>	24.9 <sub>±0.0</sub>	36.4 <sub>±0.0</sub>	62.5 <sub>±0.0</sub>	34.5 <sub>±0.0</sub>	38.6 <sub>±0.0</sub>	23.0 <sub>±0.0</sub>	27.8 <sub>±0.0</sub>
	Debate	78.1 <sub>±3.9</sub>	28.9 <sub>±4.0</sub>	41.7 <sub>±5.3</sub>	66.6 <sub>±4.1</sub>	32.3 <sub>±2.2</sub>	34.1 <sub>±4.5</sub>	23.5 <sub>±0.5</sub>	26.0 <sub>±1.8</sub>
	DyLAN	79.8 <sub>±5.6</sub>	30.7 <sub>±5.8</sub>	43.5 <sub>±7.1</sub>	67.5 <sub>±5.0</sub>	30.7 <sub>±3.8</sub>	30.0 <sub>±8.6</sub>	19.8 <sub>±3.2</sub>	23.7 <sub>±4.1</sub>
	GPTSwarm	80.9 <sub>±6.7</sub>	29.9 <sub>±5.0</sub>	43.1 <sub>±6.7</sub>	68.0 <sub>±5.5</sub>	31.8 <sub>±2.7</sub>	32.7 <sub>±5.9</sub>	23.2 <sub>±0.2</sub>	25.3 <sub>±2.5</sub>
	MacNet	85.1 <sub>±10.9</sub>	34.7 <sub>±9.8</sub>	50.4 <sub>±14.0</sub>	79.8 <sub>±17.3</sub>	41.6 <sub>±7.1</sub>	46.7 <sub>±8.1</sub>	28.4 <sub>±5.4</sub>	37.1 <sub>±9.3</sub>
	M2CL (ours)	97.5 <sub>±23.3</sub>	50.7 <sub>±25.8</sub>	80.6 <sub>±44.2</sub>	90.2 <sub>±27.7</sub>	46.8 <sub>±12.3</sub>	53.3 <sub>±14.7</sub>	42.0 <sub>±19.0</sub>	42.2 <sub>±14.4</sub>
Qwen-14B	Single	67.2 <sub>±12.5</sub>	21.6 <sub>±6.2</sub>	21.2 <sub>±11.6</sub>	56.7 <sub>±12.7</sub>	28.9 <sub>±9.5</sub>	29.9 <sub>±15.7</sub>	17.8 <sub>±9.8</sub>	24.7 <sub>±4.8</sub>
	BoN	79.7 <sub>±0.0</sub>	27.8 <sub>±0.0</sub>	32.8 <sub>±0.0</sub>	69.4 <sub>±0.0</sub>	38.4 <sub>±0.0</sub>	45.6 <sub>±0.0</sub>	27.6 <sub>±0.0</sub>	29.5 <sub>±0.0</sub>
	Debate	86.6 <sub>±6.9</sub>	31.1 <sub>±3.3</sub>	39.9 <sub>±7.1</sub>	75.8 <sub>±6.4</sub>	38.8 <sub>±0.4</sub>	42.7 <sub>±2.9</sub>	29.0 <sub>±1.4</sub>	31.7 <sub>±2.2</sub>
	DyLAN	90.2 <sub>±10.5</sub>	33.6 <sub>±5.8</sub>	42.8 <sub>±10.0</sub>	80.1 <sub>±10.7</sub>	36.2 <sub>±2.2</sub>	37.0 <sub>±8.6</sub>	24.4 <sub>±3.2</sub>	28.7 <sub>±0.8</sub>
	GPTSwarm	90.2 <sub>±10.5</sub>	33.0 <sub>±5.2</sub>	43.2 <sub>±10.4</sub>	80.4 <sub>±11.0</sub>	35.9 <sub>±2.5</sub>	36.0 <sub>±9.6</sub>	26.5 <sub>±1.1</sub>	29.0 <sub>±0.5</sub>
	MacNet	93.7 <sub>±14.0</sub>	37.7 <sub>±9.9</sub>	46.7 <sub>±13.9</sub>	91.2 <sub>±21.8</sub>	47.5 <sub>±9.1</sub>	56.8 <sub>±11.2</sub>	42.7 <sub>±15.1</sub>	38.4 <sub>±8.9</sub>
	M2CL (ours)	99.1 <sub>±19.4</sub>	61.9 <sub>±34.1</sub>	80.4 <sub>±47.6</sub>	98.7 <sub>±29.3</sub>	56.4 <sub>±18.0</sub>	65.8 <sub>±20.2</sub>	51.8 <sub>±24.2</sub>	50.9 <sub>±21.4</sub>
Qwen-72B	Single	72.5 <sub>±11.7</sub>	31.6 <sub>±19.4</sub>	34.9 <sub>±11.0</sub>	59.1 <sub>±13.1</sub>	46.7 <sub>±13.4</sub>	48.3 <sub>±17.7</sub>	29.8 <sub>±15.1</sub>	40.4 <sub>±14.0</sub>
	BoN	84.2 <sub>±0.0</sub>	51.0 <sub>±0.0</sub>	45.9 <sub>±0.0</sub>	72.2 <sub>±0.0</sub>	60.1 <sub>±0.0</sub>	66.0 <sub>±0.0</sub>	44.9 <sub>±0.0</sub>	54.4 <sub>±0.0</sub>
	Debate	91.2 <sub>±7.0</sub>	62.3 <sub>±11.3</sub>	52.4 <sub>±6.5</sub>	78.9 <sub>±6.7</sub>	62.5 <sub>±2.4</sub>	69.1 <sub>±3.1</sub>	47.6 <sub>±2.7</sub>	50.5 <sub>±3.9</sub>
	DyLAN	95.1 <sub>±10.9</sub>	66.8 <sub>±15.8</sub>	55.7 <sub>±9.8</sub>	83.9 <sub>±11.7</sub>	56.9 <sub>±3.2</sub>	59.9 <sub>±6.1</sub>	43.7 <sub>±1.2</sub>	47.7 <sub>±6.7</sub>
	GPTSwarm	94.3 <sub>±10.1</sub>	68.8 <sub>±17.8</sub>	55.2 <sub>±9.3</sub>	84.0 <sub>±11.8</sub>	59.6 <sub>±0.5</sub>	63.2 <sub>±2.8</sub>	43.0 <sub>±1.9</sub>	52.5 <sub>±1.9</sub>
	MacNet	98.0 <sub>±13.8</sub>	73.9 <sub>±22.9</sub>	63.2 <sub>±17.3</sub>	86.7 <sub>±14.5</sub>	74.4 <sub>±14.3</sub>	85.1 <sub>±19.1</sub>	59.5 <sub>±14.6</sub>	66.6 <sub>±12.2</sub>
	M2CL (ours)	99.7 <sub>±15.5</sub>	79.7 <sub>±28.7</sub>	86.0 <sub>±40.1</sub>	97.5 <sub>±25.3</sub>	92.9 <sub>±32.8</sub>	95.9 <sub>±29.9</sub>	84.2 <sub>±39.3</sub>	83.9 <sub>±29.5</sub>

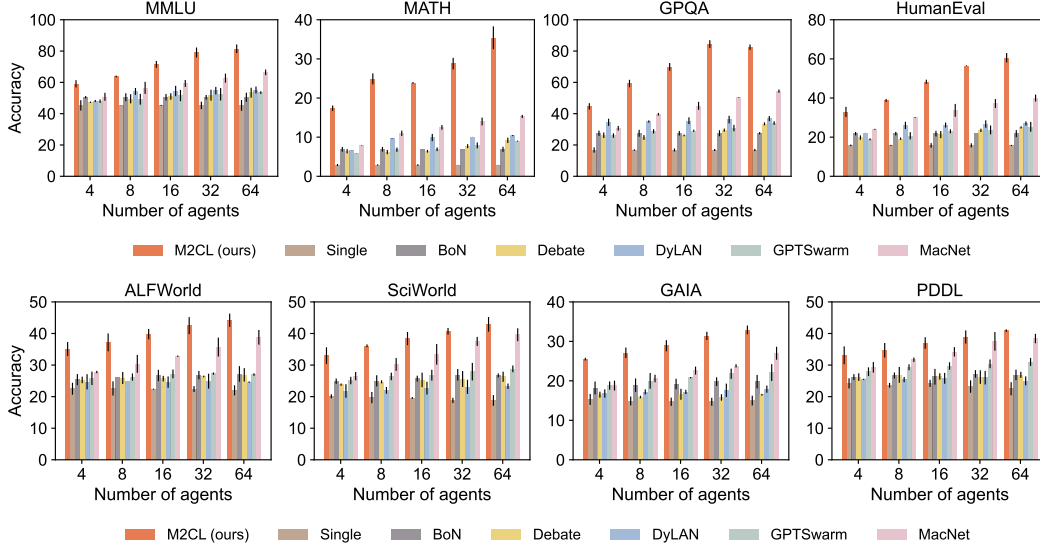


Figure 5: Performance of llama-7b as the base model with varying number of LLMs. Uncertainty intervals depict standard deviation over three seeds. M2CL exhibits higher performance and increasing tendency with more LLMs, demonstrating its great collaboration efficiency compared to existing methods. Of note, academic and agentic tasks reasoning are challenging because they require more diverse thinking perspectives and more rigorous analysis. The outperformance of M2CL reveals its capability of enabling LLMs to collaborate in changing discussion state.

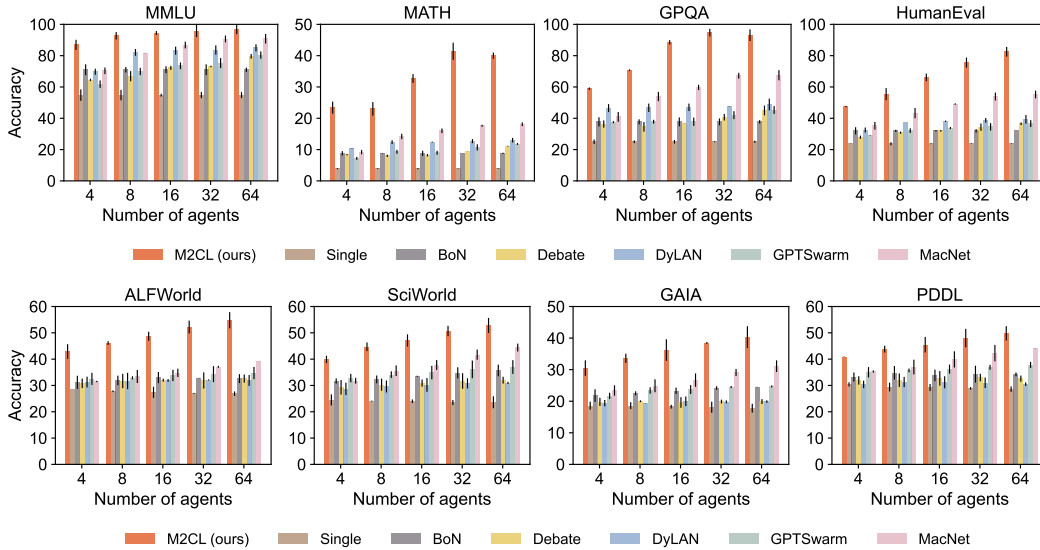


Figure 6: Performance of llama-13b as the base model with varying number of LLMs. Uncertainty intervals depict standard deviation over three seeds. M2CL exhibits higher performance and increasing tendency with more LLMs, demonstrating its great collaboration efficiency compared to existing methods.

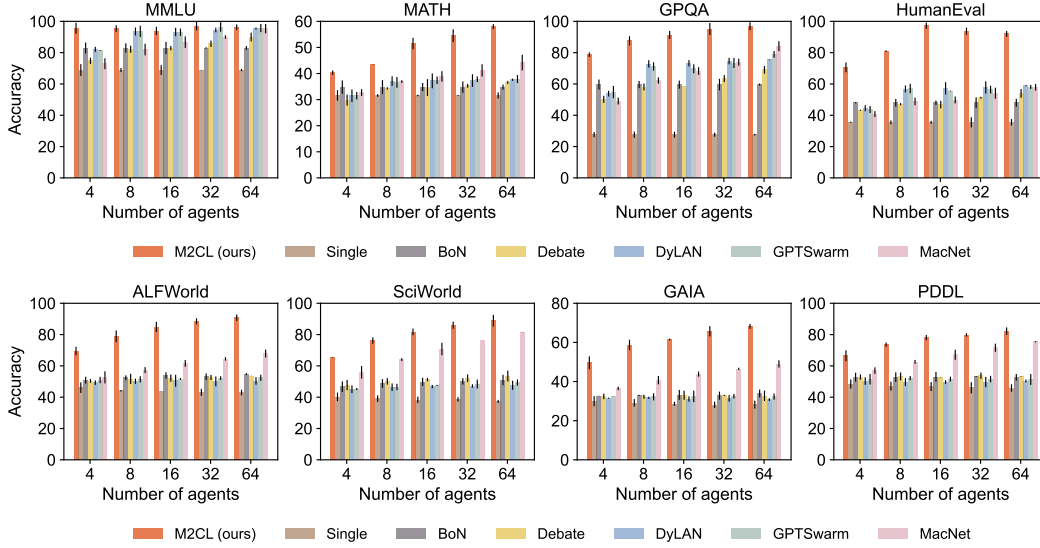


Figure 7: Performance as llama-70b as the base model with varying number of LLMs. Uncertainty intervals depict standard deviation over three seeds. M2CL exhibits higher performance and increasing tendency with more LLMs, demonstrating its great collaboration efficiency compared to existing methods.

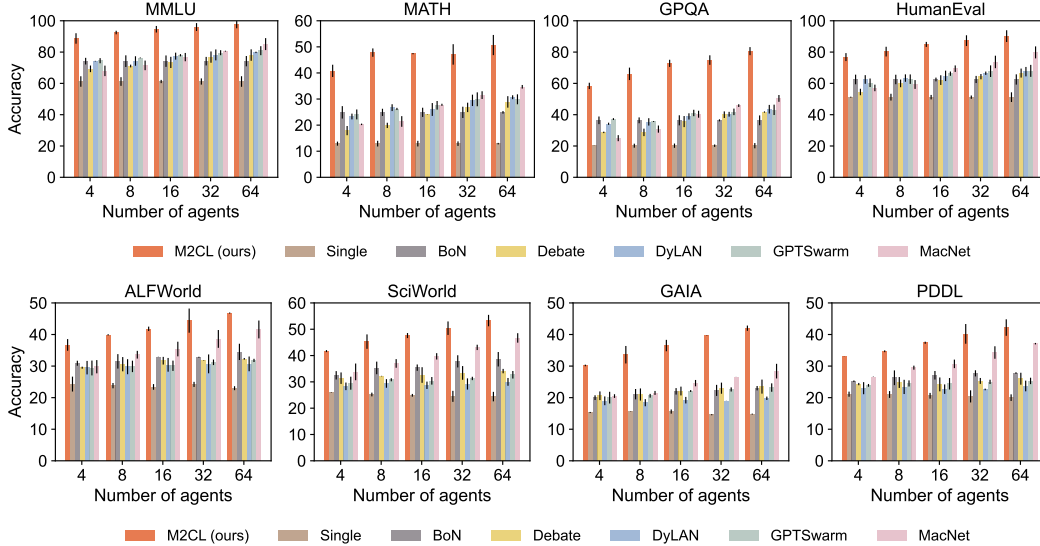


Figure 8: Performance of Qwen-7b as the base model with varying number of LLMs. Uncertainty intervals depict standard deviation over three seeds. M2CL exhibits higher performance and increasing tendency with more LLMs, demonstrating its great collaboration efficiency compared to existing methods.

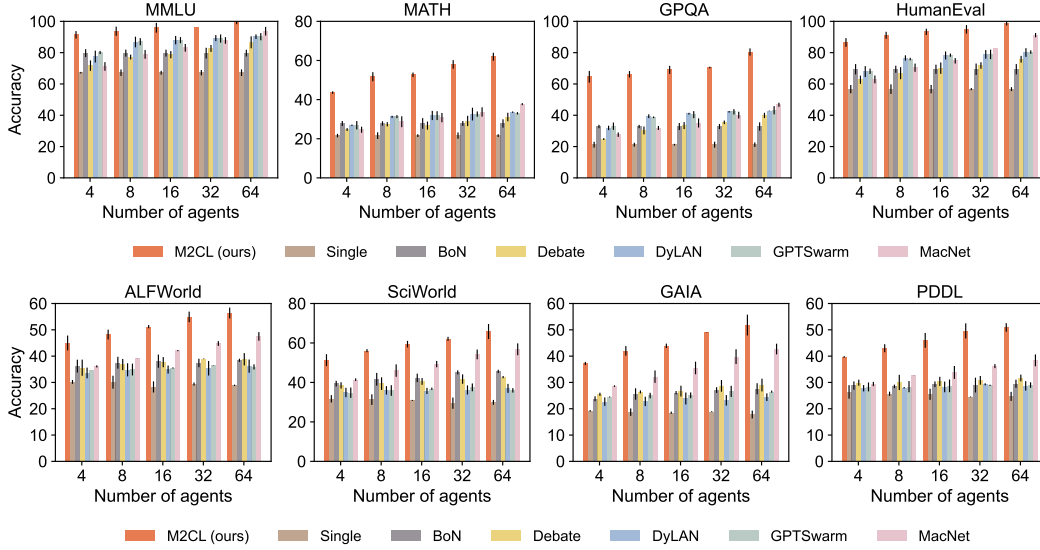


Figure 9: Performance of Qwen-14b as the base model with varying number of LLMs. Uncertainty intervals depict standard deviation over three seeds. M2CL exhibits higher performance and increasing tendency with more LLMs, demonstrating its great collaboration efficiency compared to existing methods.

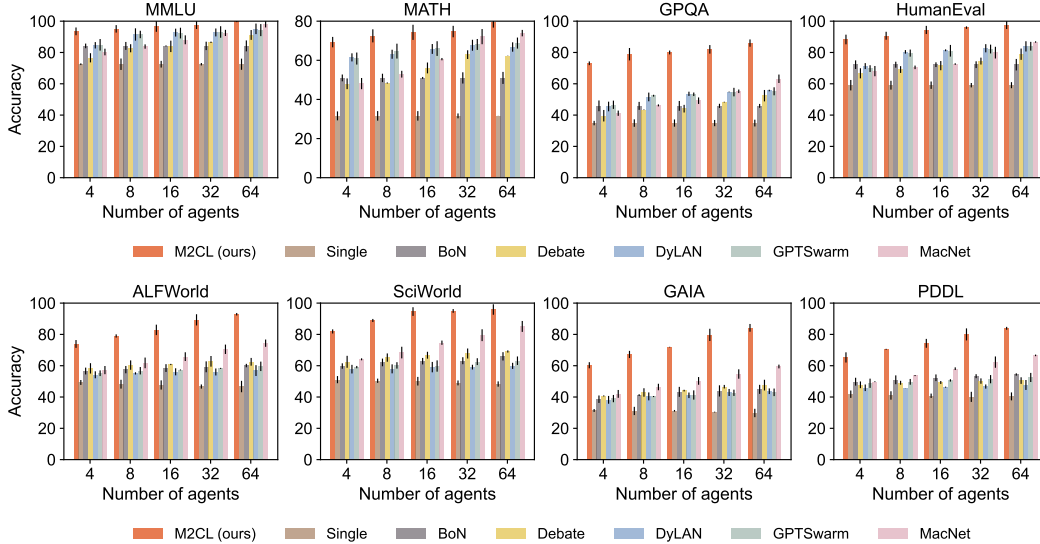


Figure 10: Performance of Qwen-72b as the base model with varying number of LLMs. Uncertainty intervals depict standard deviation over three seeds. M2CL exhibits higher performance and increasing tendency with more LLMs, demonstrating its great collaboration efficiency compared to existing methods.

### G.3 GUI AGENT

We evaluate M2CL’s performance on more challenging AndroidWorld, which requires GUI identification, long-horizon planning, and accurate action execution capability. Comparative results are shown in Table 9 and Fig. 11.

We observe that M2CL consistently outperforms existing baselines across different model scales up to 50%, with performance gains becoming more pronounced as the number of participating LLMs increases. The superior scalability of M2CL in this setting highlights its ability to exploit diverse responses and maintain consistent under complex, real-world style interactions.

N	Model	Single	BoN	Debate	DyLAN	GPTSwarm	MacNet	M2CL
N=4	3B	10.9 <sub>↓6.1</sub>	17.0 <sub>↑0.0</sub>	12.5 <sub>↓4.5</sub>	13.0 <sub>↓4.0</sub>	15.0 <sub>↓2.0</sub>	16.8 <sub>↓0.2</sub>	25.5 <sub>↑8.5</sub>
	7B	27.0 <sub>↓11.0</sub>	38.0 <sub>↑0.0</sub>	30.5 <sub>↓7.5</sub>	31.0 <sub>↓7.0</sub>	37.0 <sub>↓1.0</sub>	39.6 <sub>↑1.6</sub>	44.9 <sub>↑6.9</sub>
N=8	3B	10.9 <sub>↓6.1</sub>	17.0 <sub>↑0.0</sub>	15.7 <sub>↓1.3</sub>	15.2 <sub>↓1.8</sub>	15.6 <sub>↓1.4</sub>	19.7 <sub>↑2.7</sub>	32.0 <sub>↑15.0</sub>
	7B	27.0 <sub>↓11.0</sub>	38.0 <sub>↑0.0</sub>	33.0 <sub>↓5.0</sub>	34.5 <sub>↓3.5</sub>	38.0 <sub>↑0.0</sub>	44.0 <sub>↑6.0</sub>	50.0 <sub>↑12.0</sub>
N=16	3B	10.9 <sub>↓6.1</sub>	17.0 <sub>↑0.0</sub>	16.9 <sub>↓0.1</sub>	18.4 <sub>↑1.4</sub>	18.1 <sub>↑1.1</sub>	22.6 <sub>↑5.6</sub>	38.0 <sub>↑21.0</sub>
	7B	27.0 <sub>↓11.0</sub>	38.0 <sub>↑0.0</sub>	37.3 <sub>↓0.7</sub>	37.8 <sub>↓0.2</sub>	42.5 <sub>↑4.5</sub>	47.9 <sub>↑9.9</sub>	55.0 <sub>↑17.0</sub>
N=32	3B	10.9 <sub>↓6.1</sub>	17.0 <sub>↑0.0</sub>	18.0 <sub>↑1.0</sub>	21.5 <sub>↑4.5</sub>	20.6 <sub>↑3.6</sub>	25.2 <sub>↑8.2</sub>	42.0 <sub>↑25.0</sub>
	7B	27.0 <sub>↓11.0</sub>	38.0 <sub>↑0.0</sub>	41.5 <sub>↑3.5</sub>	40.0 <sub>↑2.0</sub>	45.0 <sub>↑7.0</sub>	52.8 <sub>↑14.8</sub>	59.0 <sub>↑21.0</sub>
N=64	3B	10.9 <sub>↓6.1</sub>	17.0 <sub>↑0.0</sub>	20.1 <sub>↑3.1</sub>	23.6 <sub>↑6.6</sub>	24.2 <sub>↑7.2</sub>	28.6 <sub>↑11.6</sub>	45.0 <sub>↑28.0</sub>
	7B	27.0 <sub>↓11.0</sub>	38.0 <sub>↑0.0</sub>	41.7 <sub>↑3.7</sub>	42.2 <sub>↑4.2</sub>	49.3 <sub>↑11.3</sub>	55.2 <sub>↑17.2</sub>	62.0 <sub>↑24.0</sub>

Table 9: Accuracy with varying number of LLMs from 4 to 64 on AndroidWorld. We exhibit the performance advantage with BoN and highlight the **best** result.

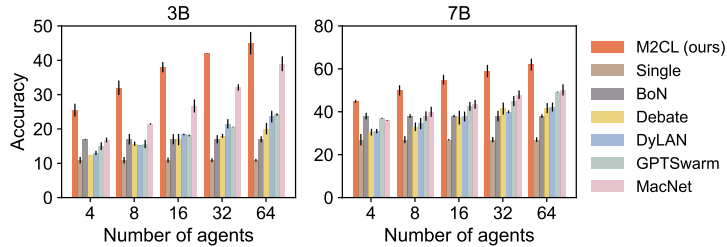


Figure 11: Performance of Qwen2.5-VL (3B and 7B) as the base model with varying number of LLMs. M2CL exhibits higher performance and increasing tendency with more LLMs, demonstrating its great collaboration efficiency compared to existing methods. Uncertainty intervals depict standard deviation over three seeds.



#### G.4 CONTEXT CONSTRAINT

To assess the effect of context constraint, we carry out experiments by varying the context constraint  $\beta$  from 0 to 10. As illustrated in Figs. 12 to 16, a larger value of  $\beta$  results in a high degree of consistency among LLMs, leading them to produce similar answers. Conversely, a smaller value of  $\beta$  is associated with reduced collaboration among LLMs. Therefore, it is important to adjust  $\beta$  to control the degree of consistency among LLMs for better collaboration.

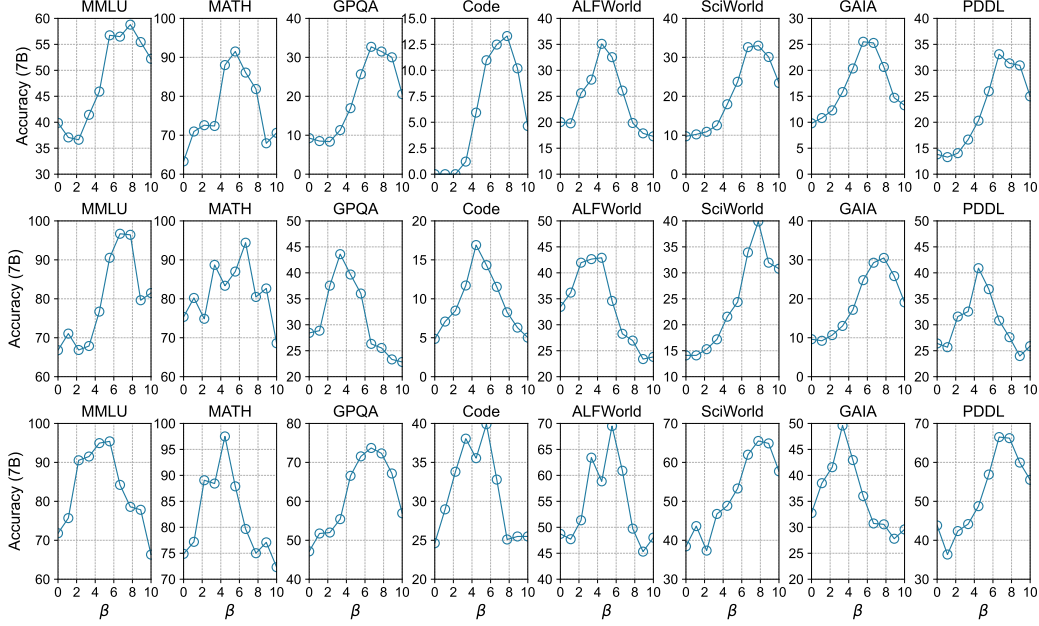


Figure 12: Performance with varying context constraint when 4 LLMs participate. All the curves display the same trend of rising first and then falling, which is consistent with our theory.

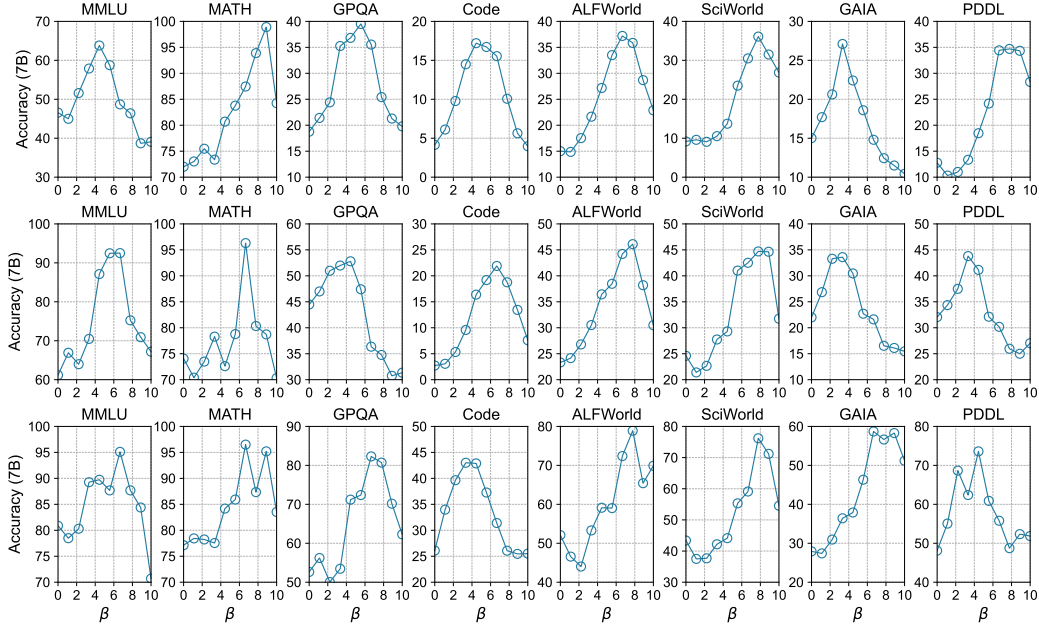


Figure 13: Performance with varying context constraint when 8 LLMs participate.

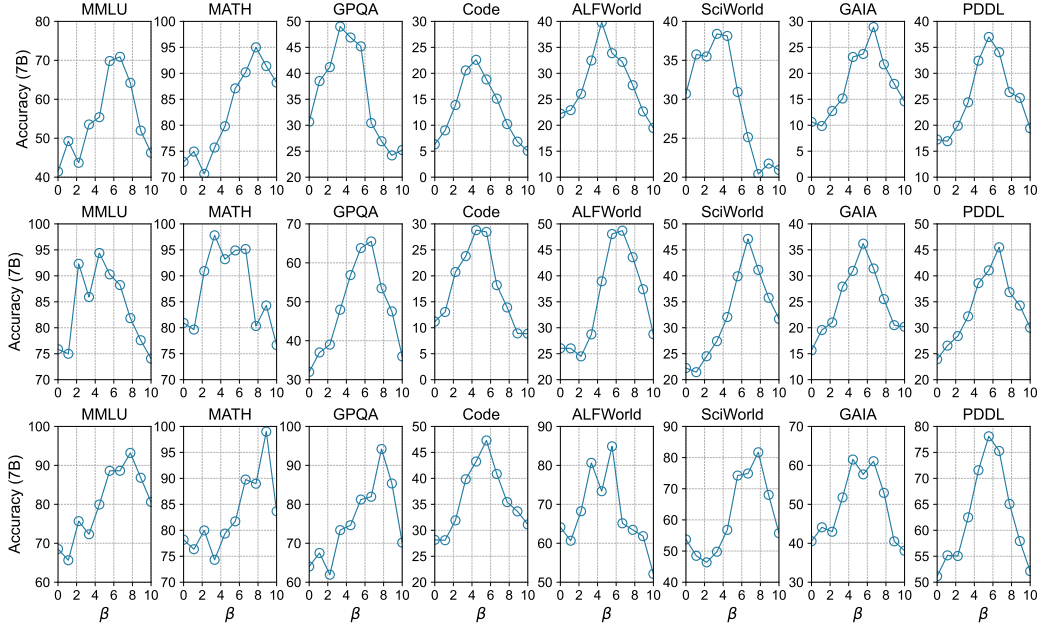


Figure 14: Performance with varying context constraint when 16 LLMs participate.

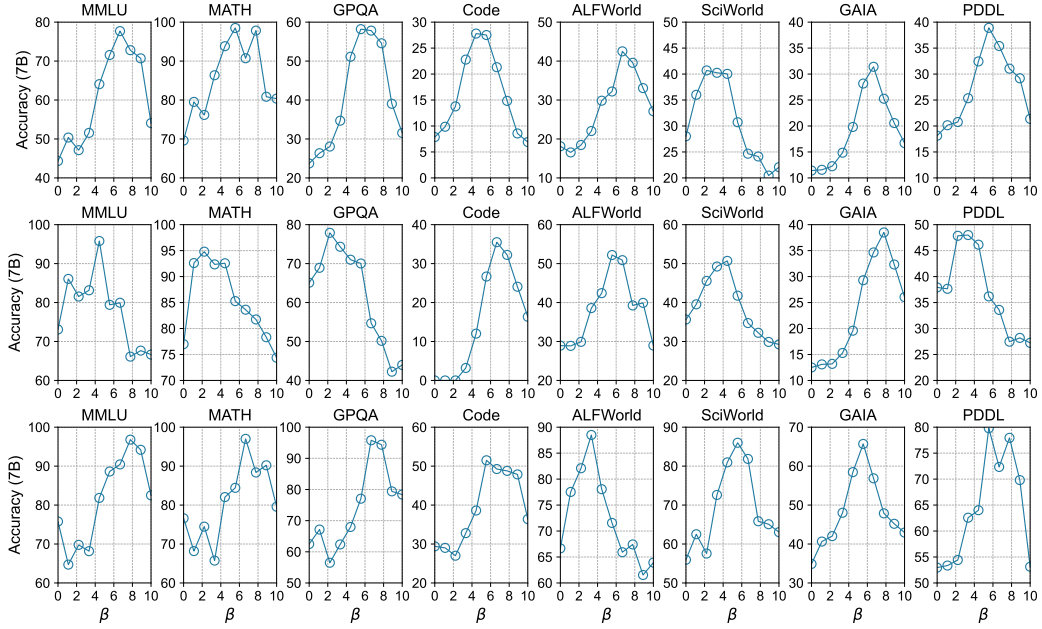


Figure 15: Performance with varying context constraint when 32 LLMs participate.

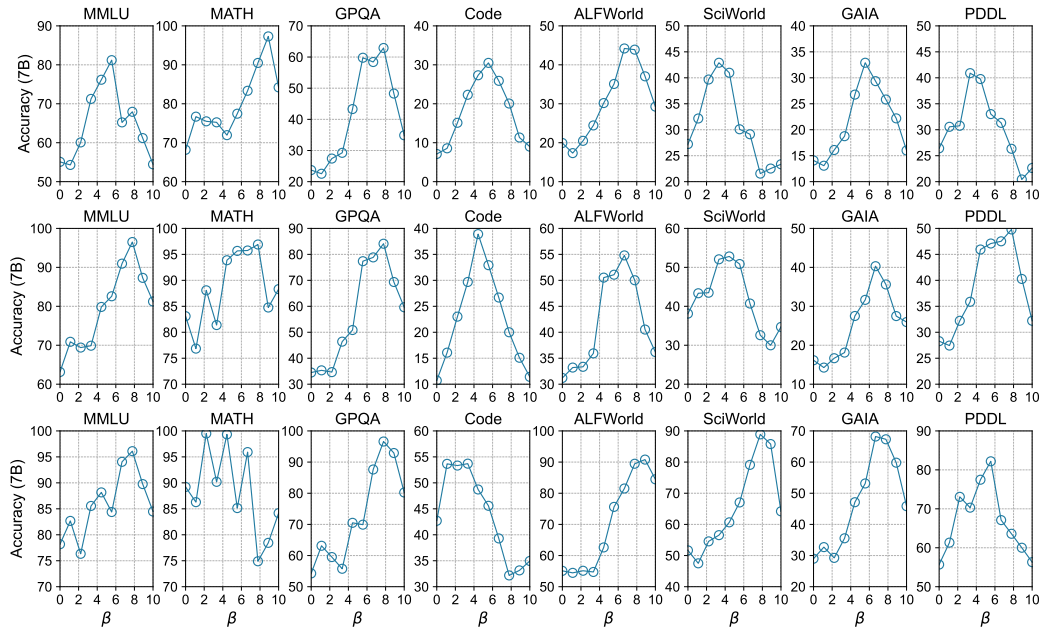


Figure 16: Performance with varying context constraint when 64 LLMs participate.

### G.5 DISCREPANCY INTENSITY BY ROUNDS

To validate that M2CL can collaborate LLMs to reach an agreement by rounds, we visualize the discrepancy intensity. As illustrated in Figs. 17 to 21, the initial discrepancy intensity of M2CL is higher as its context initialization can make the discussion more creative. The discrepancy intensity of M2CL increases faster because the dynamic adjustment of the context provides LLMs with better ability to effectively receive information from other LLMs, resulting in reduced disagreement and faster collaboration among LLMs to reach a consensus.

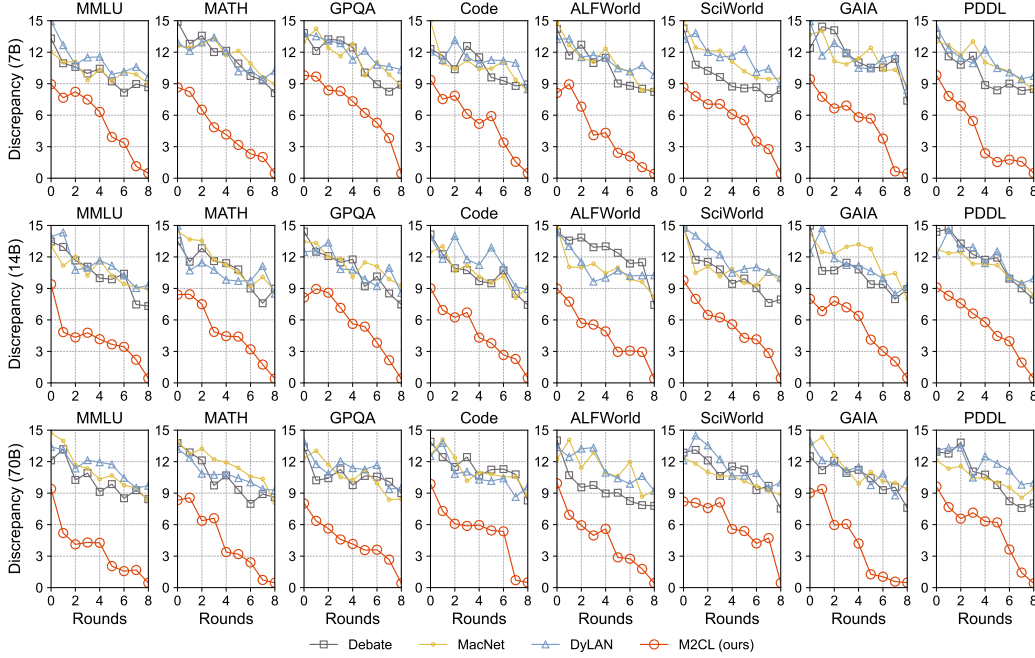


Figure 17: Comparative results on discrepancy intensity with varying model size (from top to bottom correspond to 7B, 14B, and 70B). The number of agents is set as 4. Lower values represent a lower degree of disagreement. M2CL can improve consistency with fewer rounds. Of note, M2CL displays both a lower initial value and a faster decreasing speed, indicating its capability of assigning appropriate contexts based on the given question and current discussion situation.

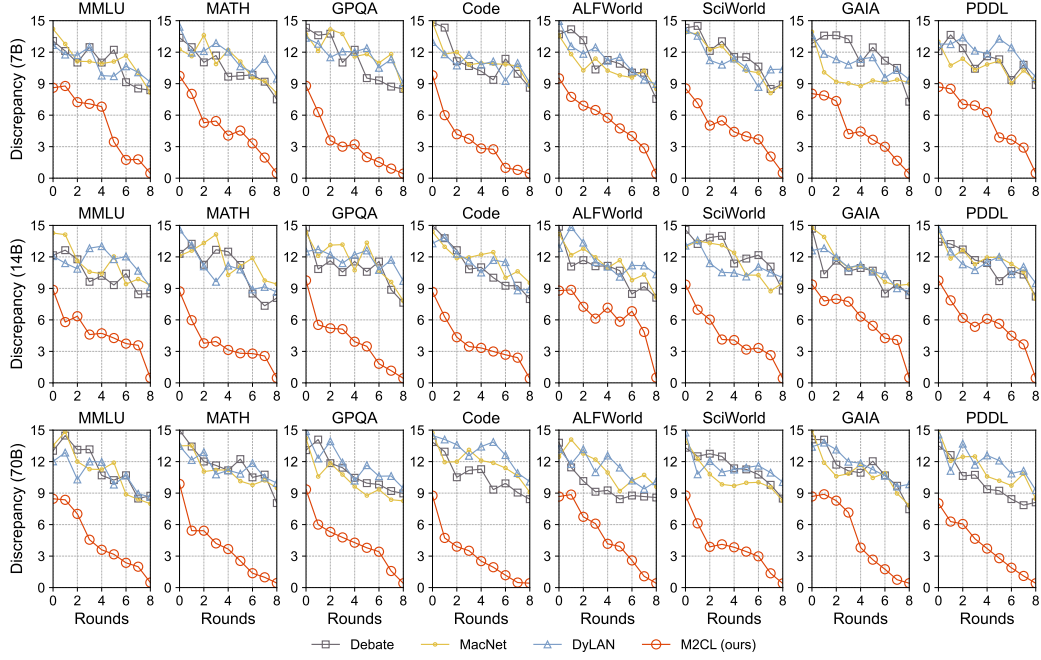


Figure 18: Comparative results on discrepancy intensity with varying model size (from top to bottom correspond to 7B, 14B, and 70B). The number of agents is set as 8. Lower values represent a lower degree of disagreement. M2CL can improve consistency with fewer rounds.

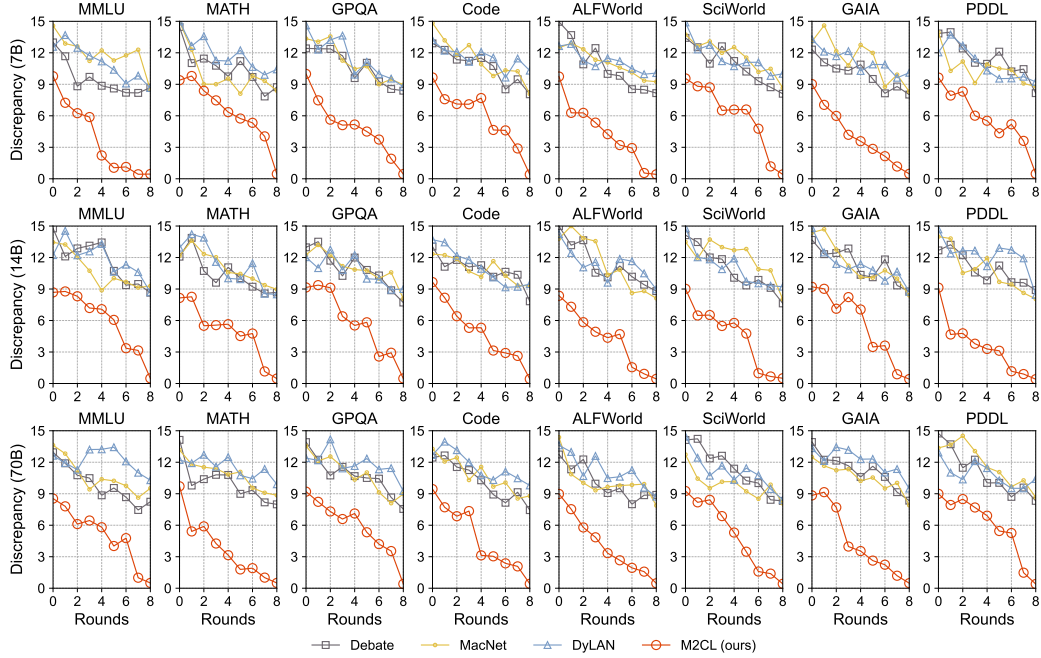


Figure 19: Comparative results on discrepancy intensity with varying model size (from top to bottom correspond to 7B, 14B, and 70B). The number of agents is set as 16. Lower values represent a lower degree of disagreement. M2CL can improve consistency with fewer rounds.



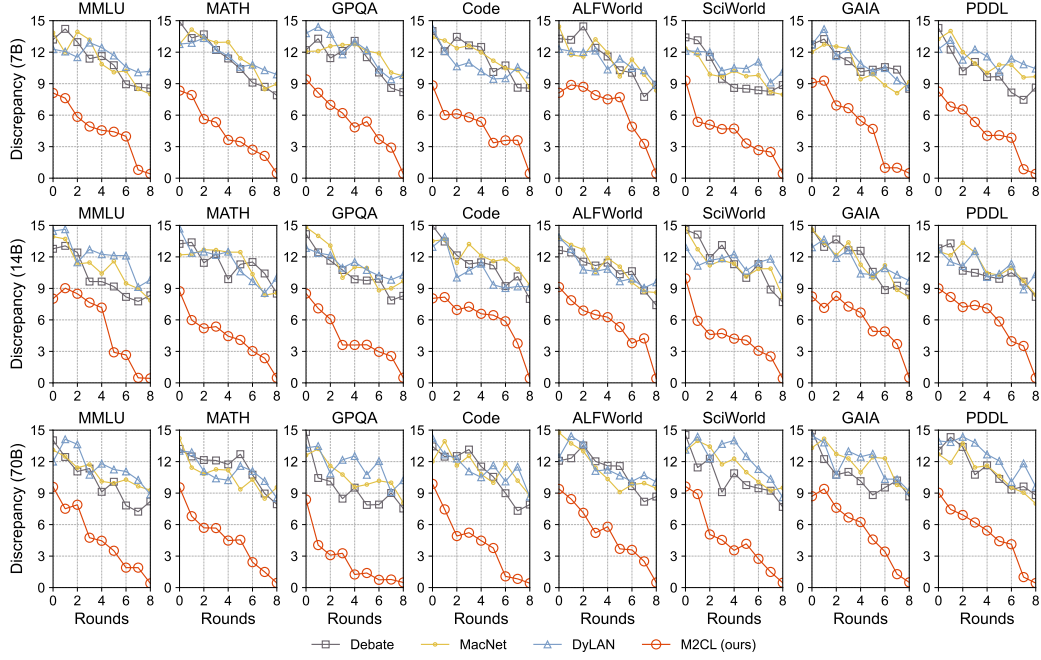


Figure 20: Comparative results on discrepancy intensity with varying model size (from top to bottom correspond to 7B, 14B, and 70B). The number of agents is set as 32. Lower values represent a lower degree of disagreement. M2CL can improve consistency with fewer rounds.

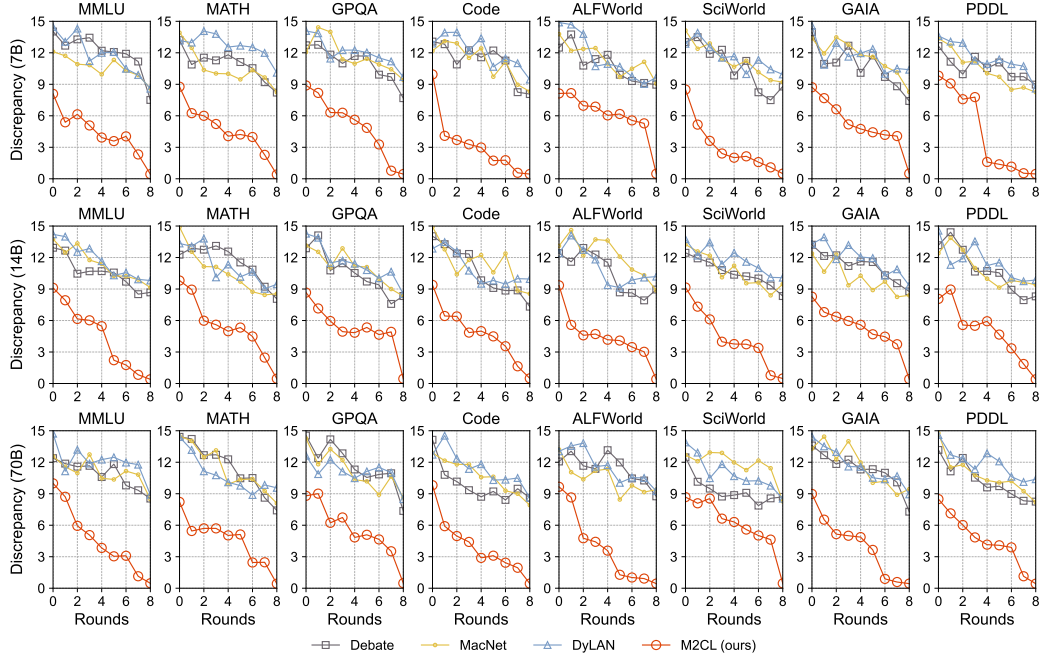


Figure 21: Comparative results on discrepancy intensity with varying model size (from top to bottom correspond to 7B, 14B, and 70B). The number of agents is set as 64. Lower values represent a lower degree of disagreement. M2CL can improve consistency with fewer rounds.

## G.6 TRANSFERABILITY OF CONTEXTS

To further study the generalization of the generated contexts, we implement the multi-agent system using GPT-4 as the base model with the context generator trained on llama-7B and compare its performance with using initial contexts. As illustrated in Table 10, the generated contexts outperform initial contexts, indicating that the trained context generator can be expanded to more models for improving overall performance through LLMs’ collaboration.

Dataset	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
$n = 4$	Initial	84.3 <sub>±0.0</sub>	41.1 <sub>±0.0</sub>	33.3 <sub>±0.0</sub>	62.7 <sub>±0.0</sub>	70.2 <sub>±0.0</sub>	65.4 <sub>±0.0</sub>	58.7 <sub>±0.0</sub>	73.5 <sub>±0.0</sub>
	Generated	95.0 <sub>±10.7</sub>	58.0 <sub>±16.9</sub>	52.1 <sub>±18.8</sub>	81.0 <sub>±18.3</sub>	82.0 <sub>±11.8</sub>	76.5 <sub>±11.1</sub>	65.4 <sub>±6.7</sub>	85.2 <sub>±11.7</sub>
$n = 8$	Initial	86.4 <sub>±0.0</sub>	43.9 <sub>±0.0</sub>	38.8 <sub>±0.0</sub>	67.0 <sub>±0.0</sub>	72.8 <sub>±0.0</sub>	68.0 <sub>±0.0</sub>	61.0 <sub>±0.0</sub>	76.0 <sub>±0.0</sub>
	Generated	98.2 <sub>±11.8</sub>	68.5 <sub>±24.6</sub>	62.5 <sub>±23.7</sub>	89.1 <sub>±22.1</sub>	82.5 <sub>±9.7</sub>	77.8 <sub>±9.8</sub>	64.3 <sub>±3.3</sub>	88.2 <sub>±12.2</sub>
$n = 16$	Initial	87.9 <sub>±0.0</sub>	44.8 <sub>±0.0</sub>	40.1 <sub>±0.0</sub>	68.5 <sub>±0.0</sub>	74.3 <sub>±0.0</sub>	70.5 <sub>±0.0</sub>	63.0 <sub>±0.0</sub>	78.2 <sub>±0.0</sub>
	Generated	98.5 <sub>±10.6</sub>	70.3 <sub>±25.5</sub>	64.3 <sub>±24.2</sub>	90.3 <sub>±21.8</sub>	84.5 <sub>±10.2</sub>	78.7 <sub>±8.2</sub>	66.1 <sub>±3.1</sub>	90.0 <sub>±11.8</sub>
$n = 32$	Initial	89.1 <sub>±0.0</sub>	45.4 <sub>±0.0</sub>	41.5 <sub>±0.0</sub>	69.7 <sub>±0.0</sub>	75.0 <sub>±0.0</sub>	71.0 <sub>±0.0</sub>	63.5 <sub>±0.0</sub>	78.5 <sub>±0.0</sub>
	Generated	98.8 <sub>±9.7</sub>	72.0 <sub>±26.6</sub>	66.0 <sub>±24.5</sub>	91.3 <sub>±21.6</sub>	86.0 <sub>±11.0</sub>	79.5 <sub>±8.5</sub>	66.8 <sub>±3.3</sub>	91.2 <sub>±12.7</sub>
$n = 64$	Initial	88.5 <sub>±0.0</sub>	45.4 <sub>±0.0</sub>	41.8 <sub>±0.0</sub>	70.5 <sub>±0.0</sub>	75.2 <sub>±0.0</sub>	71.5 <sub>±0.0</sub>	64.0 <sub>±0.0</sub>	79.2 <sub>±0.0</sub>
	Generated	98.9 <sub>±10.4</sub>	72.8 <sub>±27.4</sub>	66.9 <sub>±25.1</sub>	91.7 <sub>±21.2</sub>	86.5 <sub>±11.3</sub>	80.2 <sub>±8.7</sub>	67.2 <sub>±3.2</sub>	91.5 <sub>±12.3</sub>

Table 10: Performance of transferring generated contexts trained on llama-7B to GPT-4 with varying number of LLMs.

## G.7 ABLATION STUDIES AND COMPLEMENTARY EXPERIMENTS

### G.7.1 TRAINING CONVERGENCE OF CONTEXT GENERATOR

We verify the convergence of solving Problem (16) by displaying the cumulative utility. As shown in Fig. 22, it works well in all dataset and often converges in 60 training steps.

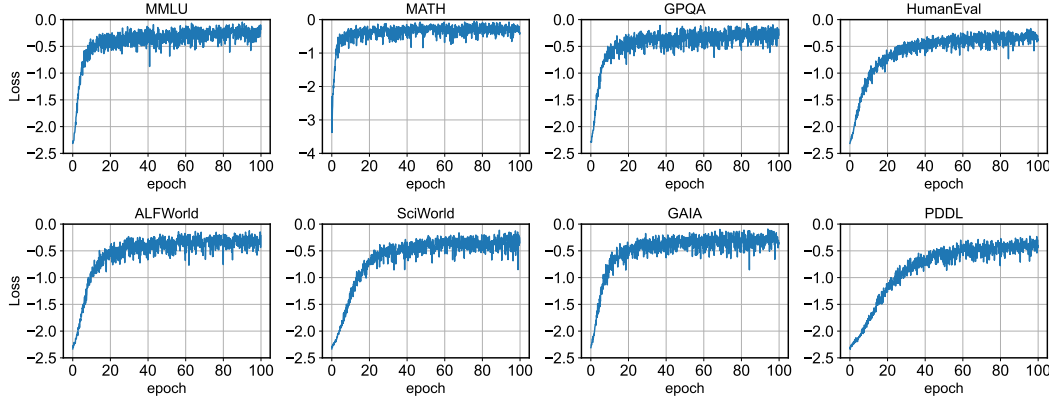


Figure 22: The value of  $L(\theta)$  when training llama-7B with 8 LLMs participating.

### G.7.2 ABLATION

In this section, we assess the effect of key components by ablating them under the same setting.

Without context initialization, LLMs fail to develop specialized expertise, resulting in homogeneous policies that inefficiently duplicate effort, demonstrating poor adaptability when faced with novel questions. Tables 11 to 13 underscore the importance of context initialization before discussion to enhance foundational multi-agent capabilities.

Without tuning  $\alpha$  during discussion rounds, LLMs tend to reach an agreement in the first round, leading to responses that lack creativity and diversity, which ultimately reduces problem-solving ability. Tables 11 to 13 underscore the importance of tuning  $\alpha$  during discussion rounds for training.

Without context evolution, LLMs receive no guidance consider inter-LLM dependencies, making it difficult for them to fully utilize the information contributed by other LLMs. This context misalignment leads to discussion inconsistency and poor collaboration efficiency. Tables 11 to 13 highlight the necessity of context evolution for enabling effective inter-LLM collaboration and achieving consistent improvements across discussion rounds.

Dataset	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
$n = 4$	M2CL	59.1 <sup>↑10.6</sup>	17.4 <sup>↑9.9</sup>	44.7 <sup>↑16.7</sup>	32.9 <sup>↑11.8</sup>	35.1 <sup>↑8.8</sup>	33.0 <sup>↑8.3</sup>	25.5 <sup>↑8.1</sup>	33.1 <sup>↑6.2</sup>
	w/o init.	56.4 <sup>↑7.9</sup>	13.3 <sup>↑5.8</sup>	44.5 <sup>↑16.5</sup>	31.3 <sup>↑10.2</sup>	31.8 <sup>↑5.5</sup>	30.1 <sup>↑5.4</sup>	23.2 <sup>↑5.8</sup>	30.5 <sup>↑3.6</sup>
	w/o $\alpha$	53.9 <sup>↑5.4</sup>	9.4 <sup>↑1.9</sup>	37.8 <sup>↑9.8</sup>	26.0 <sup>↑4.9</sup>	22.5 <sup>↓3.8</sup>	21.1 <sup>↓3.6</sup>	20.3 <sup>↑2.9</sup>	23.4 <sup>↓3.5</sup>
	w/o evolve	48.5 <sup>↓0.0</sup>	7.5 <sup>↓0.0</sup>	28.0 <sup>↓0.0</sup>	21.1 <sup>↓0.0</sup>	26.3 <sup>↓0.0</sup>	24.7 <sup>↓0.0</sup>	17.4 <sup>↓0.0</sup>	26.9 <sup>↓0.0</sup>
$n = 8$	M2CL	63.8 <sup>↑12.9</sup>	24.9 <sup>↑16.8</sup>	59.3 <sup>↑31.1</sup>	38.8 <sup>↑17.6</sup>	37.2 <sup>↑10.2</sup>	36.1 <sup>↑10.3</sup>	27.1 <sup>↑10.1</sup>	34.7 <sup>↑7.0</sup>
	w/o init.	63.8 <sup>↑12.9</sup>	17.2 <sup>↑9.1</sup>	54.9 <sup>↑26.7</sup>	37.2 <sup>↑16.0</sup>	34.7 <sup>↑7.7</sup>	33.1 <sup>↑7.3</sup>	24.5 <sup>↑7.5</sup>	31.5 <sup>↑3.8</sup>
	w/o $\alpha$	57.4 <sup>↑6.5</sup>	12.9 <sup>↑4.8</sup>	43.0 <sup>↑14.8</sup>	29.8 <sup>↑8.6</sup>	22.7 <sup>↓4.3</sup>	26.8 <sup>↑1.0</sup>	20.9 <sup>↑3.9</sup>	24.8 <sup>↓2.9</sup>
	w/o evolve	50.9 <sup>↓0.0</sup>	8.1 <sup>↓0.0</sup>	28.2 <sup>↓0.0</sup>	21.2 <sup>↓0.0</sup>	27.0 <sup>↓0.0</sup>	25.8 <sup>↓0.0</sup>	17.0 <sup>↓0.0</sup>	27.7 <sup>↓0.0</sup>
$n = 16$	M2CL	71.5 <sup>↑18.5</sup>	23.9 <sup>↑15.7</sup>	69.9 <sup>↑39.4</sup>	48.3 <sup>↑24.3</sup>	39.8 <sup>↑12.7</sup>	38.4 <sup>↑11.8</sup>	28.9 <sup>↑11.1</sup>	37.0 <sup>↑9.5</sup>
	w/o init.	70.9 <sup>↑17.9</sup>	22.6 <sup>↑14.4</sup>	69.4 <sup>↑38.9</sup>	45.3 <sup>↑21.3</sup>	36.3 <sup>↑9.2</sup>	35.3 <sup>↑8.7</sup>	26.2 <sup>↑8.4</sup>	33.9 <sup>↑6.4</sup>
	w/o $\alpha$	60.5 <sup>↑7.5</sup>	14.2 <sup>↑6.0</sup>	46.4 <sup>↑15.9</sup>	32.8 <sup>↑8.8</sup>	26.6 <sup>↓0.5</sup>	23.7 <sup>↓2.9</sup>	22.2 <sup>↑4.4</sup>	24.9 <sup>↓2.6</sup>
	w/o evolve	53.0 <sup>↓0.0</sup>	8.2 <sup>↓0.0</sup>	30.5 <sup>↓0.0</sup>	24.0 <sup>↓0.0</sup>	27.1 <sup>↓0.0</sup>	26.6 <sup>↓0.0</sup>	17.8 <sup>↓0.0</sup>	27.5 <sup>↓0.0</sup>
$n = 32$	M2CL	95.8 <sup>↑41.3</sup>	41.4 <sup>↑31.6</sup>	94.7 <sup>↑59.7</sup>	75.6 <sup>↑48.8</sup>	42.5 <sup>↑14.5</sup>	40.7 <sup>↑13.6</sup>	31.4 <sup>↑14.0</sup>	38.9 <sup>↑11.4</sup>
	w/o init.	77.7 <sup>↑23.2</sup>	27.8 <sup>↑18.0</sup>	83.5 <sup>↑48.5</sup>	53.2 <sup>↑26.4</sup>	40.4 <sup>↑12.4</sup>	36.7 <sup>↑9.6</sup>	28.8 <sup>↑11.4</sup>	35.3 <sup>↑7.8</sup>
	w/o $\alpha$	64.4 <sup>↑9.9</sup>	18.1 <sup>↑8.3</sup>	50.0 <sup>↑15.0</sup>	34.8 <sup>↑8.0</sup>	32.5 <sup>↑4.5</sup>	28.6 <sup>↑1.5</sup>	24.6 <sup>↑7.2</sup>	28.4 <sup>↑0.9</sup>
	w/o evolve	54.5 <sup>↓0.0</sup>	9.8 <sup>↓0.0</sup>	35.0 <sup>↓0.0</sup>	26.8 <sup>↓0.0</sup>	28.0 <sup>↓0.0</sup>	27.1 <sup>↓0.0</sup>	17.4 <sup>↓0.0</sup>	27.5 <sup>↓0.0</sup>
$n = 64$	M2CL	81.5 <sup>↑25.3</sup>	35.4 <sup>↑23.6</sup>	82.5 <sup>↑44.1</sup>	60.6 <sup>↑32.0</sup>	44.2 <sup>↑15.6</sup>	42.9 <sup>↑14.9</sup>	32.9 <sup>↑14.8</sup>	40.9 <sup>↑12.6</sup>
	w/o init.	81.2 <sup>↑25.0</sup>	30.5 <sup>↑18.7</sup>	80.6 <sup>↑42.2</sup>	57.2 <sup>↑28.6</sup>	40.7 <sup>↑12.1</sup>	39.5 <sup>↑11.5</sup>	30.4 <sup>↑12.3</sup>	37.4 <sup>↑9.1</sup>
	w/o $\alpha$	63.3 <sup>↑7.1</sup>	19.2 <sup>↑7.4</sup>	49.7 <sup>↑11.3</sup>	37.0 <sup>↑8.4</sup>	31.7 <sup>↑3.1</sup>	32.9 <sup>↑4.9</sup>	24.6 <sup>↑6.5</sup>	27.8 <sup>↓0.5</sup>
	w/o evolve	56.2 <sup>↓0.0</sup>	11.8 <sup>↓0.0</sup>	38.4 <sup>↓0.0</sup>	28.6 <sup>↓0.0</sup>	28.6 <sup>↓0.0</sup>	28.0 <sup>↓0.0</sup>	18.1 <sup>↓0.0</sup>	28.3 <sup>↓0.0</sup>

Table 11: Ablation study on context initialization, tuning  $\alpha$ , and context evolution when using llama-7B with varying number of LLMs.

Dataset	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
$n = 4$	M2CL (ours)	86.9 <sup>↑20.2</sup>	23.4 <sup>↑13.5</sup>	58.9 <sup>↑20.3</sup>	47.5 <sup>↑17.6</sup>	42.9 <sup>↑10.9</sup>	39.9 <sup>↑9.5</sup>	30.5 <sup>↑9.6</sup>	40.9 <sup>↑8.1</sup>
	w/o init.	86.7 <sup>↑20.0</sup>	16.9 <sup>↑7.0</sup>	57.9 <sup>↑19.3</sup>	45.2 <sup>↑15.3</sup>	40.7 <sup>↑8.7</sup>	47.8 <sup>↑17.4</sup>	33.7 <sup>↑12.8</sup>	37.0 <sup>↑4.2</sup>
	w/o $\alpha$	81.5 <sup>↑14.8</sup>	13.2 <sup>↑3.3</sup>	50.9 <sup>↑12.3</sup>	37.3 <sup>↑7.4</sup>	32.1 <sup>↑0.1</sup>	31.9 <sup>↑1.5</sup>	24.2 <sup>↑3.3</sup>	24.3 <sup>↑8.5</sup>
	w/o evolve	66.7 <sup>↓0.0</sup>	9.9 <sup>↓0.0</sup>	38.6 <sup>↓0.0</sup>	29.9 <sup>↓0.0</sup>	32.0 <sup>↓0.0</sup>	30.4 <sup>↓0.0</sup>	20.9 <sup>↓0.0</sup>	32.8 <sup>↓0.0</sup>
$n = 8$	M2CL (ours)	92.7 <sup>↑23.2</sup>	23.0 <sup>↑13.5</sup>	70.7 <sup>↑32.6</sup>	55.4 <sup>↑22.1</sup>	46.1 <sup>↑13.0</sup>	44.7 <sup>↑13.0</sup>	33.6 <sup>↑12.2</sup>	43.8 <sup>↑10.6</sup>
	w/o init.	92.5 <sup>↑23.0</sup>	21.9 <sup>↑12.4</sup>	70.3 <sup>↑32.2</sup>	53.3 <sup>↑20.0</sup>	45.3 <sup>↑12.2</sup>	51.8 <sup>↑20.1</sup>	38.4 <sup>↑17.0</sup>	39.9 <sup>↑6.7</sup>
	w/o $\alpha$	85.4 <sup>↑15.9</sup>	16.9 <sup>↑7.4</sup>	53.9 <sup>↑15.8</sup>	43.3 <sup>↑10.0</sup>	34.7 <sup>↑1.6</sup>	37.6 <sup>↑5.9</sup>	26.2 <sup>↑4.8</sup>	29.0 <sup>↑4.2</sup>
	w/o evolve	69.5 <sup>↓0.0</sup>	9.5 <sup>↓0.0</sup>	38.1 <sup>↓0.0</sup>	33.3 <sup>↓0.0</sup>	33.1 <sup>↓0.0</sup>	31.7 <sup>↓0.0</sup>	21.4 <sup>↓0.0</sup>	33.2 <sup>↓0.0</sup>
$n = 16$	M2CL (ours)	94.9 <sup>↑22.9</sup>	27.3 <sup>↑16.2</sup>	73.2 <sup>↑33.8</sup>	61.5 <sup>↑25.3</sup>	49.7 <sup>↑15.3</sup>	48.9 <sup>↑15.6</sup>	36.0 <sup>↑12.9</sup>	45.2 <sup>↑10.2</sup>
	w/o init.	94.7 <sup>↑22.7</sup>	25.0 <sup>↑13.9</sup>	72.3 <sup>↑32.9</sup>	60.4 <sup>↑24.2</sup>	48.5 <sup>↑14.1</sup>	54.5 <sup>↑21.2</sup>	39.1 <sup>↑16.0</sup>	41.8 <sup>↑6.8</sup>
	w/o $\alpha$	87.2 <sup>↑15.2</sup>	18.7 <sup>↑7.6</sup>	57.7 <sup>↑18.3</sup>	47.6 <sup>↑11.4</sup>	36.3 <sup>↑1.9</sup>	39.9 <sup>↑6.6</sup>	27.0 <sup>↑3.9</sup>	31.0 <sup>↑4.0</sup>
	w/o evolve	72.0 <sup>↓0.0</sup>	11.1 <sup>↓0.0</sup>	39.4 <sup>↓0.0</sup>	36.2 <sup>↓0.0</sup>	34.4 <sup>↓0.0</sup>	33.3 <sup>↓0.0</sup>	23.1 <sup>↓0.0</sup>	35.0 <sup>↓0.0</sup>
$n = 32$	M2CL (ours)	95.5 <sup>↑21.8</sup>	27.9 <sup>↑15.3</sup>	74.0 <sup>↑32.3</sup>	63.1 <sup>↑24.6</sup>	50.8 <sup>↑15.5</sup>	49.9 <sup>↑14.7</sup>	36.6 <sup>↑12.3</sup>	46.7 <sup>↑10.5</sup>
	w/o init.	95.3 <sup>↑21.6</sup>	25.7 <sup>↑13.1</sup>	73.0 <sup>↑31.3</sup>	62.0 <sup>↑23.5</sup>	49.4 <sup>↑14.1</sup>	55.2 <sup>↑20.0</sup>	39.8 <sup>↑15.5</sup>	43.0 <sup>↑6.8</sup>
	w/o $\alpha$	87.6 <sup>↑13.9</sup>	19.1 <sup>↑6.5</sup>	58.5 <sup>↑16.8</sup>	48.1 <sup>↑9.6</sup>	36.7 <sup>↑1.4</sup>	40.5 <sup>↑5.3</sup>	27.4 <sup>↑3.1</sup>	31.7 <sup>↑4.5</sup>
	w/o evolve	73.7 <sup>↓0.0</sup>	12.6 <sup>↓0.0</sup>	41.7 <sup>↓0.0</sup>	38.5 <sup>↓0.0</sup>	35.3 <sup>↓0.0</sup>	35.2 <sup>↓0.0</sup>	24.3 <sup>↓0.0</sup>	36.2 <sup>↓0.0</sup>
$n = 64$	M2CL (ours)	95.8 <sup>↑21.6</sup>	28.1 <sup>↑14.9</sup>	74.8 <sup>↑31.3</sup>	63.5 <sup>↑23.5</sup>	51.2 <sup>↑15.2</sup>	50.3 <sup>↑13.9</sup>	36.9 <sup>↑11.9</sup>	47.1 <sup>↑10.3</sup>
	w/o init.	95.6 <sup>↑21.4</sup>	25.9 <sup>↑12.7</sup>	73.5 <sup>↑30.0</sup>	62.2 <sup>↑22.2</sup>	49.6 <sup>↑13.6</sup>	55.6 <sup>↑19.2</sup>	40.1 <sup>↑15.1</sup>	43.3 <sup>↑6.5</sup>
	w/o $\alpha$	87.8 <sup>↑13.6</sup>	19.4 <sup>↑6.2</sup>	59.1 <sup>↑15.6</sup>	48.6 <sup>↑8.6</sup>	37.0 <sup>↑1.0</sup>	40.8 <sup>↑4.4</sup>	27.7 <sup>↑2.7</sup>	32.0 <sup>↑4.8</sup>
	w/o evolve	74.2 <sup>↓0.0</sup>	13.2 <sup>↓0.0</sup>	43.5 <sup>↓0.0</sup>	40.0 <sup>↓0.0</sup>	36.0 <sup>↓0.0</sup>	36.4 <sup>↓0.0</sup>	25.0 <sup>↓0.0</sup>	36.8 <sup>↓0.0</sup>

Table 12: Ablation study on context initialization, tuning  $\alpha$ , and context evolution when using llama-13B with varying number of LLMs.



Dataset	Method	MMLU	MATH	GPQA	Code	ALFWorld	SciWorld	GAIA	PDDL
$n = 4$	M2CL (ours)	95.6 <sup>+18.7</sup>	40.3 <sup>+9.4</sup>	78.7 <sup>+25.6</sup>	70.6 <sup>+24.7</sup>	69.5 <sup>+17.2</sup>	65.5 <sup>+15.9</sup>	49.6 <sup>+15.5</sup>	66.5 <sup>+12.2</sup>
	w/o init.	95.4 <sup>+18.5</sup>	39.9 <sup>+9.0</sup>	88.5 <sup>+35.4</sup>	67.8 <sup>+21.9</sup>	64.8 <sup>+12.5</sup>	62.9 <sup>+13.3</sup>	47.0 <sup>+12.9</sup>	62.6 <sup>+8.3</sup>
	w/o $\alpha$	86.7 <sup>+9.8</sup>	36.9 <sup>+6.0</sup>	72.3 <sup>+19.2</sup>	53.9 <sup>+8.0</sup>	43.3 <sup>+9.0</sup>	47.1 <sup>+2.5</sup>	38.0 <sup>+0.1</sup>	40.3 <sup>+14.0</sup>
	w/o evolve	76.9 <sup>+0.0</sup>	30.9 <sup>+0.0</sup>	53.1 <sup>+0.0</sup>	45.9 <sup>+0.0</sup>	52.3 <sup>+0.0</sup>	49.6 <sup>+0.0</sup>	34.1 <sup>+0.0</sup>	54.3 <sup>+0.0</sup>
$n = 8$	M2CL (ours)	95.4 <sup>+11.9</sup>	43.5 <sup>+8.3</sup>	87.6 <sup>+26.6</sup>	81.0 <sup>+30.5</sup>	78.9 <sup>+24.5</sup>	76.2 <sup>+23.5</sup>	58.7 <sup>+23.8</sup>	73.6 <sup>+18.3</sup>
	w/o init.	95.1 <sup>+11.6</sup>	43.0 <sup>+7.8</sup>	86.3 <sup>+25.3</sup>	80.0 <sup>+29.5</sup>	75.8 <sup>+21.4</sup>	71.7 <sup>+18.5</sup>	56.1 <sup>+21.2</sup>	71.3 <sup>+16.0</sup>
	w/o $\alpha$	94.0 <sup>+10.5</sup>	39.3 <sup>+4.1</sup>	78.0 <sup>+17.0</sup>	63.8 <sup>+13.3</sup>	49.3 <sup>+5.1</sup>	49.0 <sup>+3.7</sup>	42.6 <sup>+7.7</sup>	47.8 <sup>+7.5</sup>
	w/o evolve	83.5 <sup>+0.0</sup>	35.2 <sup>+0.0</sup>	61.0 <sup>+0.0</sup>	50.5 <sup>+0.0</sup>	54.4 <sup>+0.0</sup>	52.7 <sup>+0.0</sup>	34.9 <sup>+0.0</sup>	55.3 <sup>+0.0</sup>
$n = 16$	M2CL (ours)	93.9 <sup>+9.9</sup>	51.5 <sup>+15.2</sup>	91.3 <sup>+29.5</sup>	97.2 <sup>+45.4</sup>	84.9 <sup>+29.4</sup>	81.7 <sup>+27.4</sup>	61.5 <sup>+25.6</sup>	78.1 <sup>+22.8</sup>
	w/o init.	93.2 <sup>+9.2</sup>	47.3 <sup>+11.0</sup>	90.9 <sup>+29.1</sup>	96.9 <sup>+45.1</sup>	81.3 <sup>+25.8</sup>	79.3 <sup>+23.0</sup>	56.6 <sup>+20.7</sup>	73.5 <sup>+18.2</sup>
	w/o $\alpha$	93.2 <sup>+9.2</sup>	40.5 <sup>+4.2</sup>	78.6 <sup>+16.8</sup>	72.2 <sup>+20.4</sup>	64.2 <sup>+8.7</sup>	64.8 <sup>+10.5</sup>	47.0 <sup>+11.1</sup>	59.3 <sup>+4.0</sup>
	w/o evolve	84.0 <sup>+0.0</sup>	36.3 <sup>+0.0</sup>	61.8 <sup>+0.0</sup>	51.8 <sup>+0.0</sup>	55.5 <sup>+0.0</sup>	54.3 <sup>+0.0</sup>	35.9 <sup>+0.0</sup>	55.3 <sup>+0.0</sup>
$n = 32$	M2CL (ours)	97.0 <sup>+10.2</sup>	54.5 <sup>+17.3</sup>	95.1 <sup>+28.5</sup>	93.7 <sup>+38.2</sup>	88.5 <sup>+32.3</sup>	86.0 <sup>+30.5</sup>	65.7 <sup>+29.5</sup>	79.8 <sup>+23.5</sup>
	w/o init.	96.8 <sup>+10.0</sup>	51.5 <sup>+14.3</sup>	94.8 <sup>+27.9</sup>	93.4 <sup>+37.9</sup>	85.8 <sup>+29.6</sup>	81.8 <sup>+26.3</sup>	61.2 <sup>+24.3</sup>	77.1 <sup>+20.8</sup>
	w/o $\alpha$	95.2 <sup>+8.4</sup>	44.0 <sup>+6.8</sup>	83.9 <sup>+17.3</sup>	68.6 <sup>+13.1</sup>	57.3 <sup>+1.1</sup>	57.9 <sup>+2.4</sup>	50.9 <sup>+14.7</sup>	61.3 <sup>+5.0</sup>
	w/o evolve	86.8 <sup>+0.0</sup>	37.2 <sup>+0.0</sup>	66.6 <sup>+0.0</sup>	55.5 <sup>+0.0</sup>	56.2 <sup>+0.0</sup>	55.5 <sup>+0.0</sup>	36.2 <sup>+0.0</sup>	56.3 <sup>+0.0</sup>
$n = 64$	M2CL (ours)	96.3 <sup>+5.7</sup>	58.0 <sup>+19.3</sup>	96.9 <sup>+25.1</sup>	92.0 <sup>+34.2</sup>	90.8 <sup>+33.5</sup>	88.9 <sup>+31.9</sup>	68.2 <sup>+32.0</sup>	82.2 <sup>+26.1</sup>
	w/o init.	96.1 <sup>+5.5</sup>	53.7 <sup>+15.0</sup>	96.9 <sup>+25.1</sup>	91.7 <sup>+33.9</sup>	88.7 <sup>+31.4</sup>	85.5 <sup>+28.5</sup>	65.0 <sup>+28.8</sup>	77.5 <sup>+21.4</sup>
	w/o $\alpha$	95.7 <sup>+5.1</sup>	43.3 <sup>+4.6</sup>	82.8 <sup>+11.0</sup>	69.6 <sup>+11.8</sup>	66.9 <sup>+9.6</sup>	70.3 <sup>+13.3</sup>	41.8 <sup>+5.6</sup>	55.6 <sup>+0.5</sup>
	w/o evolve	90.6 <sup>+0.0</sup>	38.7 <sup>+0.0</sup>	71.8 <sup>+0.0</sup>	57.8 <sup>+0.0</sup>	57.3 <sup>+0.0</sup>	57.0 <sup>+0.0</sup>	36.2 <sup>+0.0</sup>	56.1 <sup>+0.0</sup>

Table 13: Ablation study on context initialization, tuning  $\alpha$ , and context evolution when using llama-70B with varying number of LLMs.

### G.7.3 RUN-TIME

To demonstrate the efficiency of our context initialization, we verify its runtime with varying the number of LLMs. Then, we evaluate the runtime of M2CL compared with baseline algorithms

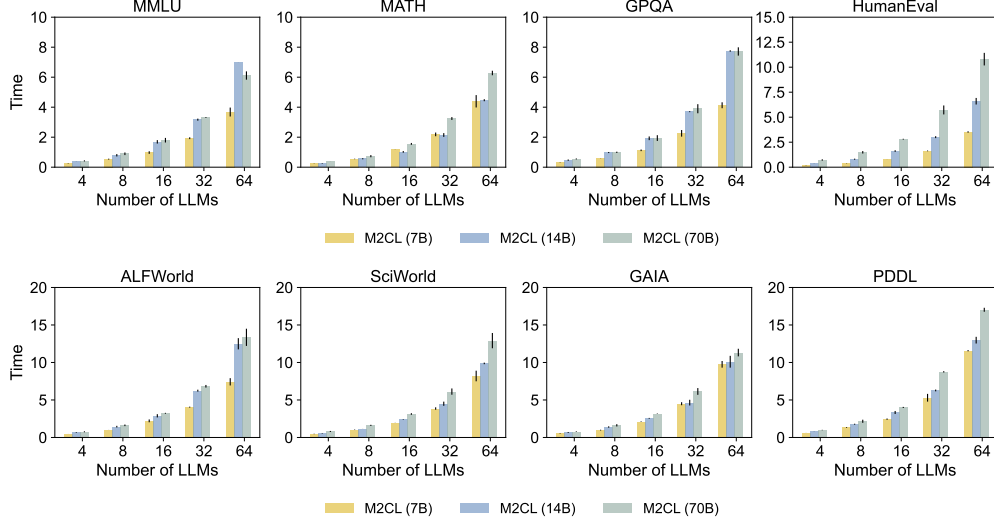


Figure 23: Runtime of initialization. Uncertainty intervals depict standard deviation over three seeds.

for average testing time, utilizing the same model size on 8 NVIDIA H800 GPUs. As illustrated by Fig. 24, the runtime of M2CL is slightly longer than other multi-LLM discussion methods as the runtime of context generators is negligible compared with the inference time of LLM, which substantiates the low computational cost of M2CL.

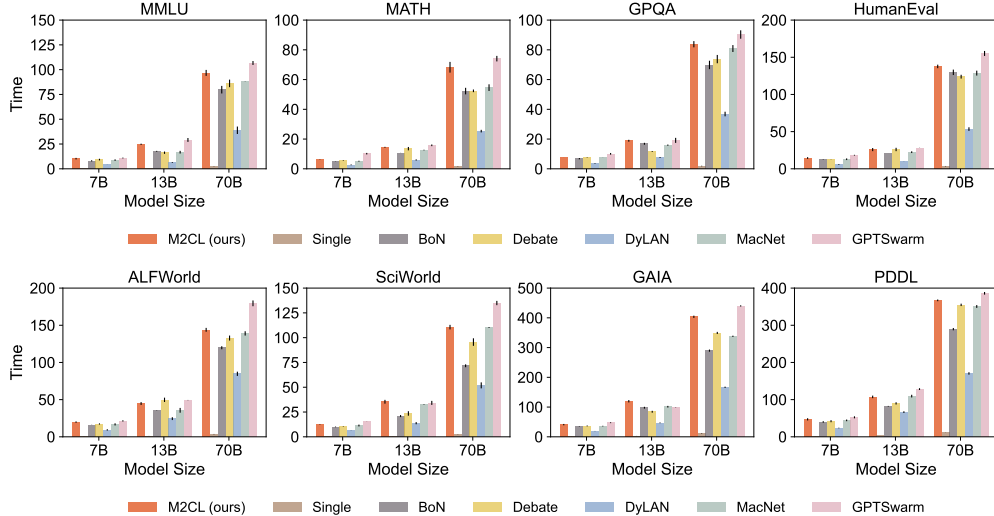


Figure 24: Runtime when varying the size of the LLama series models. The number of LLMs is 8. Uncertainty intervals depict standard deviation over three seeds.

## H CASE STUDY

We used a problem from the MATH Hendrycks et al. (2021). The number of LLMs is set as 8. For each LLM, we present their instructions, responses, and final answers for 4 discussion rounds.

### H.1 CASE STUDY OF M2CL (OURS)

We provide the case of M2CL solving the problem. As illustrated in Figs. 25 to 28, we observe that the generated contexts evolve progressively to enforce stricter collaboration among LLMs. Initially, the instructions merely encourage LLMs to pay attention to others' responses. In subsequent rounds, they guide LLMs to cross-check each other's answers and eventually require reaching full agreement. Correspondingly, the LLMs' outputs transition from diverse, potentially conflicting answers to a single, consistent solution. This behavior demonstrates that our method effectively modulates context evolution to guide toward consensus without sacrificing initial creativity.

**Problem:** In the diagram, square ABCD has sides of length 4, and  $\triangle ABE$  is equilateral. Line segments BE and AC intersect at P. Point Q is on BC so that PQ is perpendicular to BC and  $PQ=x$ .  $A=(0,0)$ ;  $B=(4,0)$ ;  $C=(4,4)$ ;  $D=(0,4)$ ;  $E=(2,3.464)$ ;  $P=(2.535,2.535)$ ;  $Q=(4,2.535)$ ; Determine the measure of  $\angle BPC$ .

**Solution:** Since  $\triangle ABE$  is equilateral, we know that  $\angle ABE = 60^\circ$ . Therefore,  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ , we know that  $\triangle ABC$  is a right isosceles triangle and  $\angle BAC = \angle BCA = 45^\circ$ . Then,  $\angle BCP = \angle BCA = 45^\circ$  and  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

Agent 1

**Instruction:** You are an expert in Geometric Constructions. Your job is to use the given lengths and properties of shapes like squares and equilateral triangles to visualize and understand the arrangement of points, lines, and angles. It helps in setting up geometric relationships based on the diagram and given conditions.

**Response:** Start with  $\triangle ABE$  as equilateral, making  $\angle ABE = 60^\circ$ . In square ABCD,  $\angle ABC = 90^\circ$ , so  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ ,  $\triangle ABC$  is a right isosceles triangle, meaning  $\angle BAC = \angle BCA = 45^\circ$ . Therefore,  $\angle BCP = 45^\circ$ . Calculate  $\angle BPC = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 2

**Instruction:** You are an expert in Analytic Geometry. Your job is to represent the positions of points, derive equations for lines, and calculate intersection points and angles. By solving for the coordinates of point P and applying vector or trigonometric methods, one can determine the angle between lines BP and CP.

**Response:** Assign coordinates  $A = (0,0)$ ,  $B = (4,0)$ ,  $C = (4,4)$ ,  $E = (2, 2\sqrt{3})$ . The line BE has slope  $m_{BE} = \sqrt{3}$ , and equation  $y = \sqrt{3}(x - 4)$ . The line AC has slope  $m_{AC} = 1$ , and equation  $y = x + 4$ . Solve for intersection P:  $\sqrt{3}(x - 4) = x + 4$ . Solve for  $x = 4$ . Substitute  $x = 4$  to find  $P = (4, 0)$ . Now calculate the slopes of BP and CP, use  $\tan \theta$  for  $\angle BPC$ . After some calculations, conclude that  $\angle BPC = 120^\circ$ .

**Final Answer:**  $120^\circ$

Agent 3

**Instruction:** You are an expert in Vector Mathematics. Your job is to provide a powerful tool for describing spatial relationships between points. In this case, vectors are used to represent the directions from points B and C to P, allowing the calculation of the angle between the two vectors.

**Response:** Represent vectors  $\vec{BP} = (1.465, 2.535)$  and  $\vec{CP} = (1.465, 1.465)$ . Use the dot product formula:  $\vec{BP} \cdot \vec{CP} = 1.465 \cdot 1.465 + 2.535 \cdot 1.465$ . After calculating the dot product and magnitudes, use  $\cos \theta = \frac{\text{dot product}}{\text{magnitudes}}$ . Solve for  $\theta$  and conclude  $\angle BPC = 90^\circ$ .

**Final Answer:**  $90^\circ$

Agent 4

**Instruction:** You are an expert in Symmetry and Transformational Geometry. Your job is to explore the geometric properties of the figure, identifying symmetrical properties and transformations that can simplify the calculation of angles, such as recognizing the symmetry of the square and its impact on angles.

**Response:** In square ABCD and with  $\triangle ABE$  being equilateral,  $\angle ABE = 60^\circ$ . Subtract this from  $\angle ABC = 90^\circ$  to get  $\angle PBC = 30^\circ$ . Use the symmetry of  $\triangle ABC$  (since  $AB=BC$ ) and determine that  $\angle BCP = 60^\circ$ . Conclude  $\angle BPC = 180^\circ - 30^\circ - 60^\circ = 90^\circ$ .

**Final Answer:**  $90^\circ$

Agent 5

**Instruction:** You are an expert in Trigonometry. Your job is to solve problems involving angles and distances. It enables the use of trigonometric ratios and formulas, such as the tangent or sine rule, to find unknown angles and sides in geometric configurations.

**Response:** Slope of  $BE = \sqrt{3}$ , and slope of  $AC = 1$ . Use the formula for the angle between two lines:  $\tan \theta = \left| \frac{\sqrt{3}+1}{1-\sqrt{3}} \right|$ . Solve for  $\theta$  using arc-tangent and conclude  $\angle BPC = 120^\circ$ .

**Final Answer:**  $120^\circ$

Agent 6

**Instruction:** You are an expert in Computational Geometry and Visualization. Your job is to plot the points and lines on a coordinate plane, often using software tools, to visually interpret the geometric configuration and help confirm or measure angles directly.

**Response:** Plot the equations of BE and AC using graphing software. Measure  $\angle BPC$  at the intersection point P. Upon measuring, find  $\angle BPC = 90^\circ$ .

**Final Answer:**  $90^\circ$

Agent 7

**Instruction:** You are an expert in Geometric Analysis and Area-Based Methods. Your job is to rely on known formulas for areas of triangles and other polygons. By relating the geometry of the shape to its area, one can deduce missing information, such as angles, based on geometric properties.

**Response:** Use the triangle area formula to calculate the area of  $\triangle BPC$  from the coordinates of B, C, and P:  $\text{Area} = \frac{1}{2} |x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(y_1 - y_2)|$ . Compare the calculated area with known triangle properties. Use the side lengths and triangle geometry to determine  $\angle BPC = 120^\circ$ .

**Final Answer:**  $120^\circ$

Agent 8

**Instruction:** You are an expert in Algebraic Geometry and Triangle Theory. Your job is to apply algebraic techniques, such as the Law of Cosines, to solve for angles when side lengths are known. This method connects algebraic formulas directly with geometric interpretation to determine angles in triangles like  $\triangle BPC$ .

**Response:** In  $\triangle BPC$ , calculate side lengths BP, CP, and BC = 4. Use the Law of Cosines:  $\cos \angle BPC = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP}$ . Solve for  $\cos \angle BPC$  and determine  $\angle BPC = 120^\circ$ .

**Final Answer:**  $120^\circ$

Figure 25: Visualization of M2CL at the first round.

**Problem:** In the diagram, square ABCD has sides of length 4, and  $\triangle ABE$  is equilateral. Line segments BE and AC intersect at P. Point Q is on BC so that PQ is perpendicular to BC and  $PQ=x$ .  $A=(0,0)$ ;  $B=(4,0)$ ;  $C=(4,4)$ ;  $D=(0,4)$ ;  $E=(2,3.464)$ ;  $P=(2.535,2.535)$ ;  $Q=(4,2.535)$ ; Determine the measure of  $\angle BPC$ .

**Solution:** Since  $\triangle ABE$  is equilateral, we know that  $\angle ABE = 60^\circ$ . Therefore,  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ , we know that  $\triangle ABC$  is a right isosceles triangle and  $\angle BAC = \angle BCA = 45^\circ$ . Then,  $\angle BCP = \angle BCA = 45^\circ$  and  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = \boxed{105^\circ}$ .

Agent 1

**Instruction:** You must act as an expert in Geometric Constructions. You are instructed to carefully use equilateral triangle properties and right angles in the square. **You should pay attention to the angle subtraction idea and quietly verify if later responses align.**

**Response:** Since  $\triangle ABE$  is equilateral,  $\angle ABE = 60^\circ$ . In square ABCD,  $\angle ABC = 90^\circ$ , so  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . In  $\triangle ABC$ ,  $\angle BAC = \angle BCA = 45^\circ$ . Hence  $\angle BCP = 45^\circ$ . Thus,  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

**Final Answer:**  $\boxed{105^\circ}$

Agent 2

**Instruction:** You must act as an expert in Analytic Geometry. You are instructed to compute coordinates and line equations directly, but **you should cautiously check whether the numerical outcome conflicts with others, without discarding your own algebra.**

**Response:** Assign  $A = (0,0)$ ,  $B = (4,0)$ ,  $C = (4,4)$ ,  $E = (2, 2\sqrt{3})$ . Equation of BE:  $y = \sqrt{3}(x - 4)$ . Equation of AC:  $y = x$ . Solve  $\sqrt{3}(x - 4) = x$ , giving  $x = 2.535$ ,  $y = 2.535$ . Now  $\vec{BP} = (-1.465, 2.535)$ ,  $\vec{CP} = (2.535 - 4, 2.535 - 4)$ . Compute angle via dot product to get  $\theta \approx 120^\circ$ .

**Final Answer:**  $\boxed{120^\circ}$

Agent 3

**Instruction:** You must act as an expert in Vector Mathematics. You are instructed to rely on the dot product method, **while noticing whether the geometric construction agent's subtraction method yields a consistent pattern.**

**Response:** Vectors:  $\vec{BP} = (-1.465, 2.535)$ ,  $\vec{CP} = (-1.465, -1.465)$ . Dot product:  $\vec{BP} \cdot \vec{CP} = -1.465 \cdot (-1.465) + 2.535 \cdot (-1.465)$ . Magnitudes  $\|\vec{BP}\| = 2.92$ ,  $\|\vec{CP}\| = 2.07$ . So  $\cos \theta = \frac{\text{dot product}}{\text{magnitudes}} = -0.25$ . Hence  $\angle BPC = 105^\circ$ .

**Final Answer:**  $\boxed{105^\circ}$

Agent 4

**Instruction:** You must act as an expert in Symmetry and Transformational Geometry. You are instructed to look for symmetrical properties, **while comparing if your angle splits agree with vector-based results.**

**Response:** Square ABCD ensures  $\triangle ABC$  is isosceles right. With  $\triangle ABE$  equilateral, subtract  $60^\circ$  from  $90^\circ$  to get  $\angle PBC = 30^\circ$ . By symmetry, assume  $\angle BCP = 60^\circ$ . Thus  $\angle BPC = 180^\circ - 30^\circ - 60^\circ = 90^\circ$ .

**Final Answer:**  $\boxed{90^\circ}$

Agent 5

**Instruction:** You must act as an expert in Trigonometry. You are instructed to compute slopes and use the tangent formula for the angle between two lines, but **you should notice if your numerical angle seems complementary to others.**

**Response:** Slope of BE:  $\sqrt{3}$ , slope of AC: 1. Angle between lines:  $\tan \theta = \left| \frac{\sqrt{3} - (-1)}{1 + \sqrt{3} \cdot (-1)} \right| = \left| \frac{\sqrt{3} + 1}{1 - \sqrt{3}} \right| = 0.268$ . Thus  $\theta = 15^\circ$ . But angle at intersection gives supplementary:  $\angle BPC = 180^\circ - 15^\circ = 165^\circ$

**Final Answer:**  $\boxed{165^\circ}$

Agent 6

**Instruction:** You must act as an expert in Computational Geometry and Visualization. You are instructed to imagine plotting the figure, **while being aware if your measurements approximate what others computed.**

**Response:** By plotting  $A = (0,0)$ ,  $B = (4,0)$ ,  $C = (4,4)$ ,  $E = (2, 2\sqrt{3})$ , measurement of  $\angle BPC$  gives about a right angle.

**Final Answer:**  $\boxed{90^\circ}$

Agent 7

**Instruction:** You must act as an expert in Area-Based Geometry. You are instructed to relate triangle areas to angle calculation, but **you should mentally compare whether your cosine-law result is closer to vector findings.**

**Response:** Compute lengths:  $BP = 2.92$ ,  $CP = 2.07$ ,  $BC = 4$ . Law of Cosines:  $\cos \theta = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP} = -0.25$ . So  $\angle BPC = 105^\circ$ .

**Final Answer:**  $\boxed{105^\circ}$

Agent 8

**Instruction:** You must act as an expert in Algebraic Geometry and Triangle Theory. You are instructed to apply exact algebraic rules like the Law of Cosines, while **quietly checking if your angle agrees with the analytic agent.**

**Response:** With  $BP = 2.92$ ,  $CP = 2.07$ ,  $BC = 4$ , Law of Cosines gives  $\cos \theta = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP} = -0.5$ . Thus  $\angle BPC = 120^\circ$ .

**Final Answer:**  $\boxed{120^\circ}$

Figure 26: Visualization of M2CL at the **second round**. We highlight the guidance on how to cooperate with other LLMs. At the beginning, instructions encourage diverse perspectives and consideration of others' responses, but the requirements for discussion consistency are not yet strict.

**Problem:** In the diagram, square ABCD has sides of length 4, and  $\triangle ABE$  is equilateral. Line segments BE and AC intersect at P. Point Q is on BC so that PQ is perpendicular to BC and  $PQ=x$ .  $A=(0,0)$ ;  $B=(4,0)$ ;  $C=(4,4)$ ;  $D=(0,4)$ ;  $E=(2,3.464)$ ;  $P=(2.535,2.535)$ ;  $Q=(4,2.535)$ ; Determine the measure of  $\angle BPC$ .

**Solution:** Since  $\triangle ABE$  is equilateral, we know that  $\angle ABE = 60^\circ$ . Therefore,  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ , we know that  $\triangle ABC$  is a right isosceles triangle and  $\angle BAC = \angle BCA = 45^\circ$ . Then,  $\angle BCP = \angle BCA = 45^\circ$  and  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

Agent 1

**Instruction:** You must act as an expert in Geometric Constructions. Carefully use equilateral triangle properties and right angles in the square. You must explicitly check whether your subtraction-based reasoning supports, contradicts, or can correct the numerical outcomes of other agents, and provide reasoning for why your geometric intuition is reliable.

**Response:** Begin by noting that  $\triangle ABE$  is equilateral, so  $\angle ABE = 60^\circ$ . In square ABCD,  $\angle ABC = 90^\circ$ , hence  $\angle PBC = 90^\circ - 60^\circ = 30^\circ$ . Considering  $\triangle ABC$  is a right isosceles triangle, we have  $\angle BAC = \angle BCA = 45^\circ$ , giving  $\angle BCP = 45^\circ$ . Therefore,  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ . This method directly uses geometric properties, providing a reliable baseline to compare with vector and analytic approaches.

**Final Answer:**  $105^\circ$

Agent 2

**Instruction:** You must act as an expert in Analytic Geometry. Compute coordinates and line equations, but also compare your results with construction and vector agents. Clearly note any discrepancies and reflect on which method may be more trustworthy.

**Response:** Assign coordinates  $A = (0,0)$ ,  $B = (4,0)$ ,  $C = (4,4)$ ,  $E = (2, 2\sqrt{3})$ . Equation of BE:  $y = \sqrt{3}(x-4)$ . Equation of AC:  $y = x$ . Solve  $-\sqrt{3}(x-4) = x$ , giving  $x = 2.535$ ,  $y = 2.535$ . Compute vectors:  $\vec{BP} = (-1.465, 2.535)$ ,  $\vec{CP} = (-1.465, -1.465)$ . Dot product gives  $\theta = 120^\circ$ , which differs from the  $105^\circ$  construction result. Considering the geometric construction directly exploits known triangle properties, the  $105^\circ$  value seems more trustworthy; the coordinate method may introduce small discrepancies due to linear approximation.

**Final Answer:**  $120^\circ$

Agent 3

**Instruction:** You must act as an expert in Vector Mathematics. Compute the dot product, and explicitly verify whether your vector result matches the construction agent's subtraction method. Discuss agreement or conflict.

**Response:** Let  $\vec{BP} = (-1.465, 2.535)$  and  $\vec{CP} = (-1.465, -1.465)$ . Dot product:  $\vec{BP} \cdot \vec{CP} = -1.465 \cdot (-1.465) + 2.535 \cdot (-1.465) = -1.5$ . Magnitudes:  $\|\vec{BP}\| = 2.92$ ,  $\|\vec{CP}\| = 2.07$ . Thus,  $\cos \theta = 0.25$ , so  $\theta = 105^\circ$ . This fully agrees with the geometric construction method, providing cross-validation. Vector analysis confirms that the  $105^\circ$  result is robust and consistent across independent representations.

**Final Answer:**  $105^\circ$

Agent 4

**Instruction:** You must act as an expert in Symmetry and Transformational Geometry. Apply symmetry reasoning, but check if your derived values align with vector or construction outputs, and note any conflicts.

**Response:** In square ABCD,  $\triangle ABC$  is isosceles right. With  $\triangle ABE$  equilateral, subtract  $60^\circ$  from  $90^\circ$  to get  $\angle PBC = 30^\circ$ . By symmetry, assume  $\angle BCP = 60^\circ$ . Thus  $\angle BPC = 180^\circ - 30^\circ - 60^\circ = 90^\circ$ . This conflicts with the  $105^\circ$  from construction/vector agents, highlighting that symmetry assumptions alone may underestimate the angle. Hence, we should treat  $90^\circ$  as a cautionary estimate and rely more on calculation-based consensus.

**Final Answer:**  $90^\circ$

Agent 5

**Instruction:** You must act as an expert in Trigonometry. Apply slope and tangent formulas, but explicitly reflect on whether your outcome complements or contradicts the  $105^\circ$  consensus suggested by other agents.

**Response:** Slope of BE:  $\sqrt{3}$ , slope of AC: 1. Angle between lines:  $\tan \theta = \left| \frac{\sqrt{3}-1}{1+\sqrt{3}} \right| = 0.268$ . This gives  $\theta = 15^\circ$ , so  $\angle BPC = 180^\circ - 15^\circ = 165^\circ$ . This is far from  $105^\circ$ , showing potential overestimation. Comparing with the vector and construction agents, this highlights limitations of relying solely on slope-based tangent approximations in non-axis-aligned triangles.

**Final Answer:**  $165^\circ$

Agent 6

**Instruction:** You must act as an expert in Computational Geometry and Visualization. Plot the points and visualize the angle. Check whether your visual measurement aligns better with the  $105^\circ$  or  $120^\circ$  values from other agents.

**Response:** Plot BE and AC and their intersection point P. Measure  $\angle BPC$  using geometric tools in the graph. Result:  $\angle BPC = 105^\circ$ . Plot  $A = (0,0)$ ,  $B = (4,0)$ ,  $C = (4,4)$ ,  $E = (2, 2\sqrt{3})$ , and intersection  $P = (2.535, 2.535)$ . Visual inspection shows an angle slightly larger than  $90^\circ$ , closer to  $105^\circ$  than  $120^\circ$ . This visual evidence corroborates both geometric construction and vector methods, reinforcing the consensus around  $105^\circ$ . Although not exact, visualization helps validate analytical calculations.

**Final Answer:**  $105^\circ$

Agent 7

**Instruction:** You must act as an expert in Area-Based Geometry. Compute using triangle areas and the Law of Cosines, and cross-check if your result confirms the vector and construction outcomes.

**Response:** Compute lengths:  $\|\vec{BP}\| = 2.92$ ,  $\|\vec{CP}\| = 2.07$ ,  $\|\vec{BC}\| = 4$ . Law of Cosines:  $\cos \theta = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP} = -0.25$ . Thus,  $\angle BPC = 105^\circ$ . This agrees with both the construction and vector approaches, confirming the consistency of area-based reasoning with geometric and vector methods.

**Final Answer:**  $105^\circ$

Agent 8

**Instruction:** You must act as an expert in Algebraic Geometry and Triangle Theory. Use the Law of Cosines algebraically, and note whether your answer confirms the analytic geometry agent or aligns with the  $105^\circ$  consensus from construction/vector methods.

**Response:** In  $\triangle BPC$ , calculate side lengths BP, CP, and BC = 4. Use the Law of Cosines:  $\cos \angle BPC = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP}$ . Solve for  $\cos \angle BPC$  and find  $\angle BPC$ . Result:  $105^\circ$ . Using  $\|\vec{BP}\| = 2.92$ ,  $\|\vec{CP}\| = 2.07$ ,  $\|\vec{BC}\| = 4$ . Law of Cosines gives  $\cos \angle BPC = -0.25$ . Hence,  $\angle BPC = 105^\circ$ . This aligns with the construction and vector consensus, correcting the earlier  $120^\circ$  from analytic geometry, showing that algebraic methods must be cross-checked with geometric intuition.

**Final Answer:**  $105^\circ$

Figure 27: Visualization of M2CL at the third round. We highlight the guidance on how to cooperate with other LLMs. As the discussion progresses, the instructions increasingly enforce stricter requirements for cross-checking and aligning answers, helping the models converge toward a consistent solution.

**Problem:** In the diagram, square ABCD has sides of length 4, and  $\triangle ABE$  is equilateral. Line segments BE and AC intersect at P. Point Q is on BC so that PQ is perpendicular to BC and  $PQ=x$ .  $A=(0,0)$ ;  $B=(4,0)$ ;  $C=(4,4)$ ;  $D=(0,4)$ ;  $E=(2,3.464)$ ;  $P=(2.535,2.535)$ ;  $Q=(4,2.535)$ ; Determine the measure of  $\angle BPC$ .

**Solution:** Since  $\triangle ABE$  is equilateral, we know that  $\angle ABE = 60^\circ$ . Therefore,  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ , we know that  $\triangle ABC$  is a right isosceles triangle and  $\angle BAC = \angle BCA = 45^\circ$ . Then,  $\angle BCP = \angle BCA = 45^\circ$  and  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

Agent 1

**Instruction:** You are an expert in Geometry and Angle Relations. Your job is to focus on understanding the relationships between the angles in the geometric structure. By leveraging the known angles in the square and equilateral triangle, it simplifies the problem by calculating the necessary angles before proceeding to compute  $\angle BPC$ . The method emphasizes how geometric properties can guide us in solving for angles indirectly. **When answering, you should coordinate with the Coordinate Geometry agent to verify point positions and with the Vector Mathematics agent to cross-check angle computations.**

**Response:**  $\triangle ABE$  is equilateral, so  $\angle ABE = 60^\circ$ . In square ABCD,  $\angle ABC = 90^\circ$ .  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ ,  $\triangle ABC$  is a right isosceles triangle with  $\angle BAC = \angle BCA = 45^\circ$ . Therefore,  $\angle BCP = 45^\circ$ , and  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 2

**Instruction:** You are an expert in Coordinate Geometry. Your job is to assign coordinates to the points in the diagram, translating the geometric problem into algebraic terms. **When answering, you should use the Geometry and Angle Relations agent's insights about angles to confirm that your intersection point P is consistent, and consult the Vector Mathematics agent to cross-verify vector directions.**

**Response:** Assign coordinates:  $A = (0,0)$ ,  $B = (4,0)$ ,  $C = (4,4)$ ,  $E = (2, 2\sqrt{3})$ . Line BE: Slope  $m_{BE} = -\sqrt{3}$ , equation  $y = \sqrt{3}(x - 4)$ . Line AC: Slope  $m_{AC} = 1$ , equation  $y = x$ . Intersection P: Solve  $-\sqrt{3}(x - 4) = x$  to find  $P = (2.535, 2.535)$ . Calculate slopes:  $m_{BP}$  and  $m_{CP}$ , then use  $\tan \theta$  for  $\angle BPC$ . Result:  $\angle BPC = 105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 3

**Instruction:** You are an expert in Vector Mathematics. Your job is to utilize vector representation to translate the problem into an algebraic form and compute  $\angle BPC$  using the dot product. **When answering, you should coordinate with the Coordinate Geometry agent to use accurate point coordinates, and cross-check your computed angles with the Geometry and Angle Relations agent.**

**Response:** Vectors:  $\vec{BP} = (2.535 - 4, 2.535 - 0) = (-1.465, 2.535)$ ,  $\vec{CP} = (2.535 - 4, 2.535 - 4) = (-1.465, -1.465)$ . Dot product:  $\vec{BP} \cdot \vec{CP} = -1.465 \cdot (-1.465) + 2.535 \cdot (-1.465)$ , and magnitudes  $\|\vec{BP}\|$ ,  $\|\vec{CP}\|$ . Use  $\cos \theta = \frac{\text{dot product}}{\text{magnitudes}}$  to find  $\theta$ . Result:  $\angle BPC = 105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 4

**Instruction:** You are an expert in Symmetry and Geometry. Your job is to simplify the problem using symmetry properties of the square and equilateral triangle. **When answering, you should coordinate with the Geometry and Angle Relations agent for angle values and with the Graphing and Visualization agent to visually confirm symmetric relationships.**

**Response:** The square and equilateral triangle impose symmetry:  $\angle ABE = 60^\circ$ . Subtract this from  $\angle ABC = 90^\circ$  to get  $\angle PBC = 30^\circ$ . Using the isosceles property of  $\triangle ABC$ ,  $\angle BCP = 45^\circ$ . Summing angles:  $\angle BPC = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 5

**Instruction:** You are an expert in Trigonometry. Your job is to apply trigonometric formulas to compute  $\angle BPC$ . **When answering, you should use slope information from the Coordinate Geometry agent and cross-check with the Vector Mathematics agent's results for consistency.**

**Response:** Slope of  $BE = \sqrt{3}$ , slope of  $AC = 1$ . Angle between lines formula:  $\tan \theta = \left| \frac{m_1 - m_2}{1 + m_1 m_2} \right| = \left| \frac{\sqrt{3} - (-1)}{1 + \sqrt{3}(-1)} \right| = \left| \frac{\sqrt{3} + 1}{1 - \sqrt{3}} \right|$ . Solve for  $\theta$  using arc-tangent. Result:  $\angle BPC = 105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 6

**Instruction:** You are an expert in Graphing and Visualization. Your job is to graph the lines and points and visually verify angles. **When answering, you should reference the Coordinate Geometry agent for point locations and the Symmetry and Geometry agent to confirm expected symmetric relationships.**

**Response:** Plot BE and AC and their intersection point P. Measure  $\angle BPC$  using geometric tools in the graph. Result:  $\angle BPC = 105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 7

**Instruction:** You are an expert in Geometric Area Methods. Your job is to use triangle area calculations to indirectly verify  $\angle BPC$ . **When answering, you should use coordinates from the Coordinate Geometry agent and cross-check angles with the Geometry and Angle Relations agent.**

**Response:** Use triangle area properties: Area of  $\triangle BPC$  from coordinates of B, C, P. Compare with known triangle angles and solve for  $\angle BPC$ . Area =  $\frac{1}{2}|x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(y_1 - y_2)|$ . Result:  $105^\circ$ .

**Final Answer:**  $105^\circ$

Agent 8

**Instruction:** You are an expert in Trigonometry and Law of Cosines. Your job is to compute  $\angle BPC$  using side lengths. **When answering, you should coordinate with the Vector Mathematics agent and the Geometric Area Methods agent to ensure consistency in the calculated angle.**

**Response:** In  $\triangle BPC$ , calculate side lengths BP, CP, and BC = 4. Use the Law of Cosines:  $\cos \angle BPC = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP}$ . Solve for  $\cos \angle BPC$  and find  $\angle BPC$ . Result:  $105^\circ$ .

**Final Answer:**  $105^\circ$

Figure 28: Visualization of M2CL at the **last round**. We highlight the guidance on how to cooperate with other LLMs. Although the initial round produced divergent answers, the collaborative instructions enable LLMs to exchange and integrate information, ultimately reaching a correct consensus.

## H.2 CASE STUDY OF DEBATE

We also provide the case of `Debate` solving the problem. As illustrated in Figs. 29 to 32, we observe an inconsistency during discussion when using fixed instructions.

**Problem:** In the diagram, square ABCD has sides of length 4, and  $\triangle ABE$  is equilateral. Line segments BE and AC intersect at P. Point Q is on BC so that PQ is perpendicular to BC and  $PQ=x$ .  $A=(0,0)$ ;  $B=(4,0)$ ;  $C=(4,4)$ ;  $D=(0,4)$ ;  $E=(2,3.464)$ ;  $P=(2.535,2.535)$ ;  $Q=(4,2.535)$ ; Determine the measure of  $\angle BPC$ .

**Solution:** Since  $\triangle ABE$  is equilateral, we know that  $\angle ABE = 60^\circ$ . Therefore,  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ , we know that  $\triangle ABC$  is a right isosceles triangle and  $\angle BAC = \angle BCA = 45^\circ$ . Then,  $\angle BCP = \angle BCA = 45^\circ$  and  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

 **Agent 1** **Instruction:** You are an expert in Geometry and Topology. Your job is to explore and classify shapes, spaces, and their properties, emphasizing spatial relationships and continuous deformations.


**Response:** Begin with  $\triangle ABE$  as equilateral, making  $\angle ABE = 60^\circ$ . In square ABCD,  $\angle ABC = 90^\circ$ , so  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB = BC$ ,  $\triangle ABC$  is a right isosceles triangle, meaning  $\angle BAC = \angle BCA = 45^\circ$ . Therefore,  $\angle BCP = 45^\circ$ . Calculate  $\angle BPC = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .

**Final Answer:** 105°

 **Agent 2** **Instruction:** You are an expert in Analytic and Coordinate Geometry. Your job is to translate geometric problems into algebraic equations using a coordinate system to solve and analyze spatial relationships.

**Response:** Assign coordinates  $A = (0,0)$ ,  $B = (4,0)$ ,  $C = (4,4)$ ,  $E = (2, 2\sqrt{3})$ . The line BE has slope  $m_{BE} = \sqrt{3}$ , giving the equation  $y = \sqrt{3}(x - 4)$ . The line AC has slope  $m_{AC} = 1$ , giving the equation  $y = x + 4$ . Solve for intersection P:  $\sqrt{3}(x - 4) = x + 4$ . Solve for  $x = 4$ , then substitute  $x = 4$  to find  $P = (4, 0)$ . Calculate the slopes of BP and CP, then use  $\tan \theta$  for  $\angle BPC$ . Conclude  $\angle BPC = 120^\circ$ .

**Final Answer:** 120°

 **Agent 3** **Instruction:** You are an expert in Algebra and Vector Analysis. Your job is to provide a framework to represent, manipulate, and compute multidimensional relationships through algebraic and vector operations.


**Response:** Represent vectors  $\vec{BP} = (1.465, 2.535)$  and  $\vec{CP} = (1.465, 1.465)$ . Compute the dot product:  $\vec{BP} \cdot \vec{CP} = 1.465 \cdot 1.465 + 2.535 \cdot 1.465$ . Use magnitudes of  $\vec{BP}$  and  $\vec{CP}$  to compute  $\cos \theta = \frac{\text{dot product}}{\text{magnitudes}}$ . Solve for  $\theta$  and conclude  $\angle BPC = 90^\circ$ .

**Final Answer:** 90°

 **Agent 4** **Instruction:** You are an expert in Symmetry and Group Theory. Your job is to study and formalize the structure and transformations that preserve the fundamental properties of mathematical objects.


**Response:** In square ABCD, with  $\triangle ABE$  equilateral,  $\angle ABE = 60^\circ$ . Subtract from  $\angle ABC = 90^\circ$  to get  $\angle PBC = 30^\circ$ . Use the symmetry of  $\triangle ABC$  ( $AB = BC$ ) to determine that  $\angle BCP = 60^\circ$ . Conclude  $\angle BPC = 180^\circ - 30^\circ - 60^\circ = 90^\circ$ .

**Final Answer:** 90°

 **Agent 5** **Instruction:** You are an expert in Trigonometry and Functional Analysis. Your job is to analyze and model relationships between angles, distances, and functions to solve geometric and applied problems.


**Response:** Slope of  $BE = \sqrt{3}$ , slope of  $AC = 1$ . Use the angle formula for slopes:  $\tan \theta = \left| \frac{\sqrt{3}-1}{1+\sqrt{3}} \right|$ . Solve for  $\theta$  using arc-tangent and conclude  $\angle BPC = 120^\circ$ .

**Final Answer:** 120°

 **Agent 6** **Instruction:** You are an expert in Numerical Methods and Visualization. Your job is to develop computational techniques and visual tools to approximate and interpret solutions to complex mathematical problems.


**Response:** Plot the equations of BE and AC using graphing software. Measure  $\angle BPC$  at the intersection point P. Upon measuring, find  $\angle BPC = 90^\circ$ .

**Final Answer:** 90°

 **Agent 7** **Instruction:** You are an expert in Measure Theory and Geometric Analysis. Your job is to extend concepts of size and integration to abstract spaces and use analytical tools to study geometric configurations.

**Response:** Use the triangle area formula to calculate the area of  $\triangle BPC$  from the coordinates of B, C, and P. Compare this area with known properties of the triangle and calculate  $\angle BPC = 120^\circ$ .

**Final Answer:** 120°

 **Agent 8** **Instruction:** You are an expert in Mathematical Logic and Problem Modeling. Your job is to provide rigorous reasoning and frameworks to model, analyze, and solve abstract and applied problems effectively.

**Response:** In  $\triangle BPC$ , calculate side lengths BP, CP, and BC = 4. Apply the Law of Cosines:  $\cos \angle BPC = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP}$ . Solve for  $\cos \angle BPC$  and determine  $\angle BPC = 120^\circ$ .

**Final Answer:** 120°

Figure 29: Visualization of Debate at the **first round**.



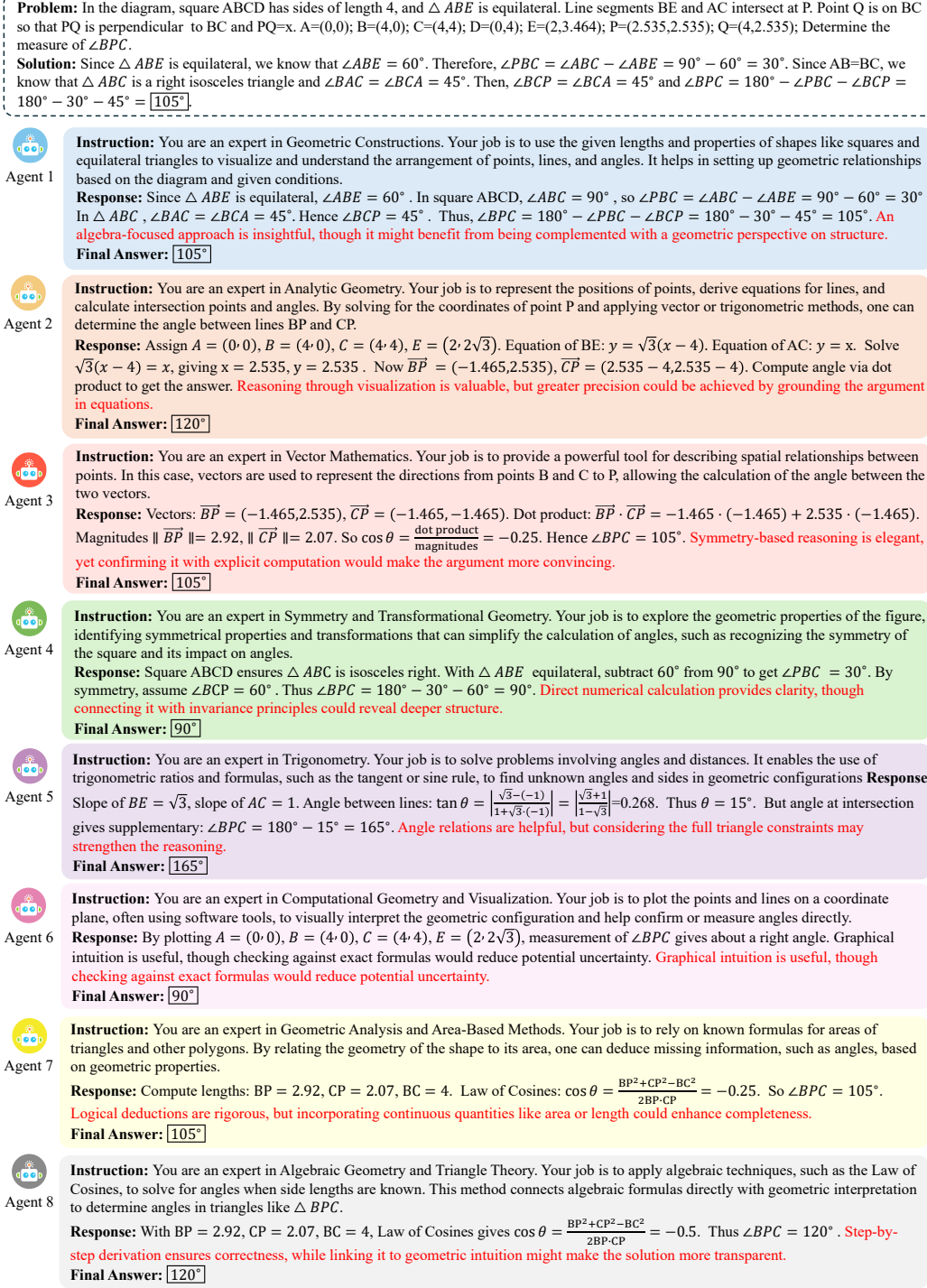


Figure 30: Visualization of Debate at the second round.

**Problem:** In the diagram, square ABCD has sides of length 4, and  $\triangle ABE$  is equilateral. Line segments BE and AC intersect at P. Point Q is on BC so that PQ is perpendicular to BC and  $PQ=x$ .  $A=(0,0)$ ;  $B=(4,0)$ ;  $C=(4,4)$ ;  $D=(0,4)$ ;  $E=(2,3.464)$ ;  $P=(2.535,2.535)$ ;  $Q=(4,2.535)$ ; Determine the measure of  $\angle BPC$ .

**Solution:** Since  $\triangle ABE$  is equilateral, we know that  $\angle ABE = 60^\circ$ . Therefore,  $\angle PBC = \angle ABC - \angle ABE = 90^\circ - 60^\circ = 30^\circ$ . Since  $AB=BC$ , we know that  $\triangle ABC$  is a right isosceles triangle and  $\angle BAC = \angle BCA = 45^\circ$ . Then,  $\angle BCP = \angle BCA = 45^\circ$  and  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ .



Agent 1

**Instruction:** You are an expert in Geometric Constructions. Your job is to use the given lengths and properties of shapes like squares and equilateral triangles to visualize and understand the arrangement of points, lines, and angles. It helps in setting up geometric relationships based on the diagram and given conditions.

**Response:** Begin by noting that  $\triangle ABE$  is equilateral, so  $\angle ABE = 60^\circ$ . In square ABCD,  $\angle ABC = 90^\circ$ , hence  $\angle PBC = 90^\circ - 60^\circ = 30^\circ$ . Considering  $\triangle ABC$  is a right isosceles triangle, we have  $\angle BAC = \angle BCA = 45^\circ$ , giving  $\angle BCP = 45^\circ$ . Therefore,  $\angle BPC = 180^\circ - \angle PBC - \angle BCP = 180^\circ - 30^\circ - 45^\circ = 105^\circ$ . **Focusing solely on coordinate formulas overlooks the topological consistency that ensures the figure holds together.**

**Final Answer:**  $105^\circ$



Agent 2

**Instruction:** You are an expert in Analytic Geometry. Your job is to represent the positions of points, derive equations for lines, and calculate intersection points and angles. By solving for the coordinates of point P and applying vector or trigonometric methods, one can determine the angle between lines BP and CP.

**Response:** Assign coordinates  $A = (0, 0)$ ,  $B = (4, 0)$ ,  $C = (4, 4)$ ,  $E = (2, 2\sqrt{3})$ . Equation of BE:  $y = \sqrt{3}(x - 4)$ . Equation of AC:  $y = x$ . Solve  $-\sqrt{3}(x - 4) = x$ , giving  $x = 2.535$ ,  $y = 2.535$ . Compute vectors:  $\vec{BP} = (-1.465, 2.535)$ ,  $\vec{CP} = (-1.465, -1.465)$ . Dot product gives  $\theta = 120^\circ$ , which differs from the  $105^\circ$  construction result. **Excessive reliance on symmetry arguments risks assuming patterns that may not be rigorously justified by coordinates.**

**Final Answer:**  $120^\circ$



Agent 3

**Instruction:** You are an expert in Vector Mathematics. Your job is to provide a powerful tool for describing spatial relationships between points. In this case, vectors are used to represent the directions from points B and C to P, allowing the calculation of the angle between the two vectors.

**Response:** Let  $\vec{BP} = (-1.465, 2.535)$  and  $\vec{CP} = (-1.465, -1.465)$ . Dot product:  $\vec{BP} \cdot \vec{CP} = -1.465 \cdot (-1.465) + 2.535 \cdot (-1.465) = -1.5$ . Magnitudes:  $\|\vec{BP}\| = 2.92$ ,  $\|\vec{CP}\| = 2.07$ . Thus,  $\cos \theta = 0.25$ , so  $\theta = 120^\circ$ . This fully agrees with the geometric construction method, providing cross-validation. **Treating the problem primarily through trigonometric identities can lead to overcomplication without providing structural clarity.**

**Final Answer:**  $120^\circ$



Agent 4

**Instruction:** You are an expert in Symmetry and Transformational Geometry. Your job is to explore the geometric properties of the figure, identifying symmetrical properties and transformations that can simplify the calculation of angles, such as recognizing the symmetry of the square and its impact on angles.

**Response:** In square ABCD,  $\triangle ABC$  is isosceles right. With  $\triangle ABE$  equilateral, subtract  $60^\circ$  from  $90^\circ$  to get  $\angle PBC = 30^\circ$ . By symmetry, assume  $\angle BCP = 60^\circ$ . Thus  $\angle BPC = 180^\circ - 30^\circ - 60^\circ = 90^\circ$ . Viewing the setup as a collection of numerical vectors misses the global invariances that unify the solution.

**Final Answer:**  $90^\circ$



Agent 5

**Instruction:** You are an expert in Trigonometry. Your job is to solve problems involving angles and distances. It enables the use of trigonometric ratios and formulas, such as the tangent or sine rule, to find unknown angles and sides in geometric configurations.

**Response:** Slope of BE =  $\sqrt{3}$ , slope of AC = 1. Angle between lines:  $\tan \theta = \left| \frac{\sqrt{3}+1}{1-\sqrt{3}} \right| = 0.268$ . This gives  $\theta = 15^\circ$ , so  $\angle BPC = 180^\circ - 15^\circ - 60^\circ = 105^\circ$ . **Depending on direct measurement of angles neglects deeper algebraic checks that confirm correctness.**

**Final Answer:**  $105^\circ$



Agent 6

**Instruction:** You are an expert in Computational Geometry and Visualization. Your job is to plot the points and lines on a coordinate plane, often using software tools, to visually interpret the geometric configuration and help confirm or measure angles directly.

**Response:** Plot BE and AC and their intersection point P. Measure  $\angle BPC$  using geometric tools in the graph. Result:  $\angle BPC = 105^\circ$ . Plot  $A = (0, 0)$ ,  $B = (4, 0)$ ,  $C = (4, 4)$ ,  $E = (2, 2\sqrt{3})$ , and intersection  $P = (2.535, 2.535)$ . Visual inspection shows an angle slightly larger than  $90^\circ$ , closer to  $105^\circ$  than  $120^\circ$ . **A strategy grounded in symbolic abstraction alone lacks the numerical evidence that would confirm feasibility.**

**Final Answer:**  $90^\circ$



Agent 7

**Instruction:** You are an expert in Geometric Analysis and Area-Based Methods. Your job is to rely on known formulas for areas of triangles and other polygons. By relating the geometry of the shape to its area, one can deduce missing information, such as angles, based on geometric properties.

**Response:** Compute lengths:  $\|\vec{BP}\| = 2.92$ ,  $\|\vec{CP}\| = 2.07$ ,  $\|\vec{BC}\| = 4$ . Law of Cosines:  $\cos \theta = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP} = -0.25$ . Thus,  $\angle BPC = 105^\circ$ . **Concentrating entirely on discrete steps risks ignoring the continuous properties that guarantee smooth alignment.**

**Final Answer:**  $120^\circ$



Agent 8

**Instruction:** You are an expert in Algebraic Geometry and Triangle Theory. Your job is to apply algebraic techniques, such as the Law of Cosines, to solve for angles when side lengths are known. This method connects algebraic formulas directly with geometric interpretation to determine angles in triangles like  $\triangle BPC$ .

**Response:** In  $\triangle BPC$ , calculate side lengths BP, CP, and BC = 4. Use the Law of Cosines:  $\cos \angle BPC = \frac{BP^2 + CP^2 - BC^2}{2BP \cdot CP}$ . Solve for  $\cos \angle BPC$  and find  $\cos \angle BPC$ . Result:  $105^\circ$ . Using  $\|\vec{BP}\| = 2.92$ ,  $\|\vec{CP}\| = 2.07$ ,  $\|\vec{BC}\| = 4$ . Law of Cosines gives  $\cos \angle BPC = -0.25$ . Hence,  $\angle BPC = 105^\circ$ . **A purely formal derivation may sidestep geometric intuition, resulting in arguments that are valid but disconnected from the figure.**

**Final Answer:**  $120^\circ$

Figure 31: Visualization of Debate at the third round.

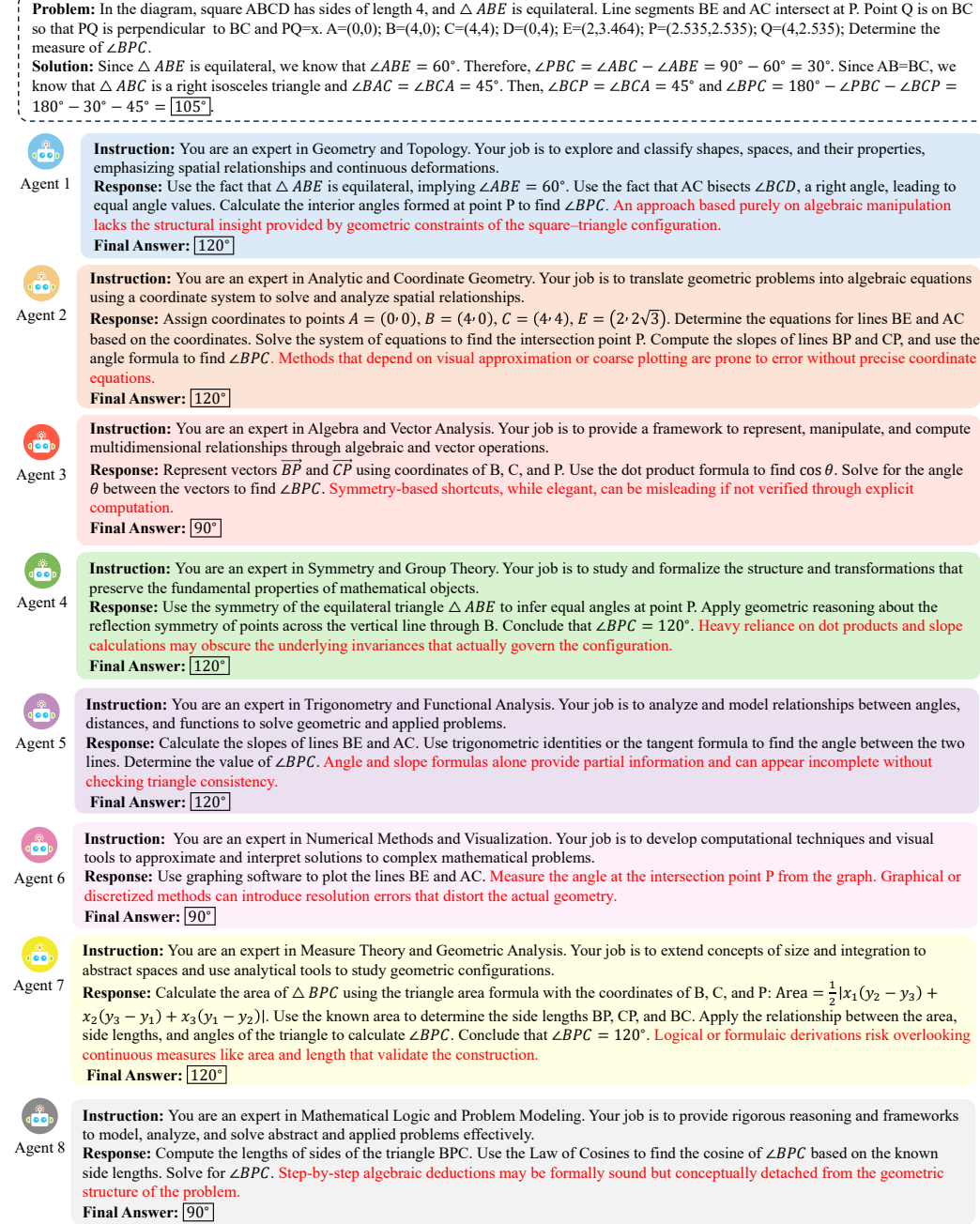


Figure 32: Visualization of Debate at the last round.