Analyze–Compose–Execute: A Dynamic Dialogue Framework for Multi-Agent Debate

Wenyuan Gu[1] *, **Haowen Wang**[2] *, **Jiale Han**[3†], **Xiang Li**[1], **Zhixuan Wu**[1], **Hongru Xiao**[4], **Bo Cheng**[1] †

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[2]School of Computer Science and Technology, Anhui University
[3]Hong Kong University of Science and Technology
[4]College of Civil Engineering, Tongji University
guwenyuan@bupt.edu.cn, wanghaowen@ahu.edu.cn, jialehan@ust.hk, lixiang2022@bupt.edu.cn, wzxmogu@bupt.edu.cn, hongru_xiao@tongji.edu.cn, chengbo@bupt.edu.cn

Abstract

Multi-Agent Debate (MAD) is an emerging paradigm that leverages the reasoning abilities of Large Language Models (LLMs) by encouraging them to collaboratively solve problems through human-like discussions. However, current MAD methods typically constrain agents to follow fixed discussion pipelines, repeatedly applying the same discussion act for a predetermined number of rounds, which limits their effectiveness and adaptability in complex and diverse tasks. To address this limitation, we propose Analyze–Compose–Execute (ACE), a novel debate framework in which agents dynamically execute the discussion actions according to the dialogue context. By analyzing the current responses of agents, ACE selects appropriate acts from a predefined Atomic Discussion Acts Library (ADAL), which are composed into a discussion action to be executed in the next round, to enable truly dynamic debate. We conduct extensive experiments on the challenging benchmark Big-Bench Hard (BBH) benchmark. ACE achieves state-of-the-art results on 17 out of 23 tasks, with an average performance gain of 8.5% across all tasks, demonstrating the effectiveness and robustness of our approach.

Introduction

Large language models (LLMs) (Brown et al. 2020a; Touvron et al. 2023; Touvron and Martin 2023; Achiam et al. 2023; Chowdhery et al. 2023) have demonstrated impressive capabilities across a wide range of natural language understanding and reasoning(Sun et al. 2021; Wang et al. 2023a; Chawla et al. 2023). Numerous studies have explored collaboration among multiple LLMs to enhance performance (Zhang et al. 2024a; Lu et al. 2025; Cai et al. 2025), with Multi-Agent Debate (MAD)(Li et al. 2023) emerging as a particularly prominent paradigm. In MAD, multiple LLM-based agents participate in the discussion process by iteratively revising their responses by incorporating insights from peer agents. Through several rounds of interaction, agents refine their reasoning and converge on more robust outputs (Tran et al. 2025; Hua et al. 2023).

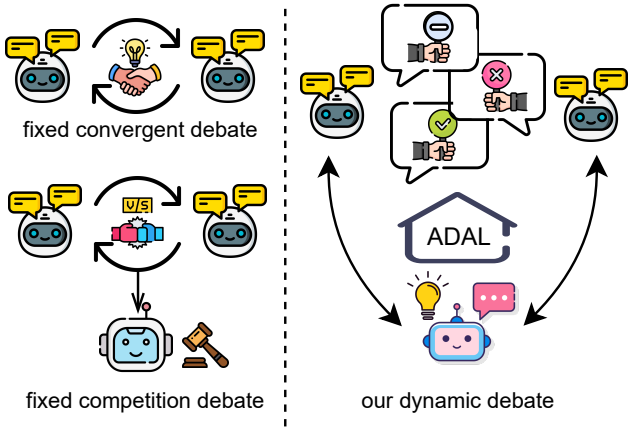

Figure 1: Comparison between existing fixed MAD methods and our proposed dynamic debate method.

Through an in-depth analysis of MAD approaches (Park et al. 2022; Hong et al. 2024; Chen et al. 2023; Gu et al. 2024), we categorize existing methods into two main paradigms. (1) *Convergent debate*, in which agents revise their responses by incorporating information or reasoning from other agents, such as multi-role discussions, roundtable meetings(Chan et al. 2023; Chen, Saha, and Bansal 2023; Zhang et al. 2024b; Li et al. 2024). (2) *Competition debate*, where agents adhere to their initial responses while engaging in mutual critique of opposing viewpoints, such as refutation-based discussion(Liang et al. 2023; Fan et al. 2024). As shown in Figure 1. Despite variations in discussion formats, these methods uniformly follow a predetermined, fixed discussion pipeline. Specifically, current MAD methods (Wang et al. 2025; Siyu Li and Zhao 2023) typically predefine a single discussion action for agents, such as repeatedly refuting the answers of each other, and require agents to perform this action throughout the entire discussion. Such fixed discussion procedures restrict agents from

*These authors contributed equally.
†Both are Corresponding authors.

adapting to complex and dynamic tasks. Under convergent debate, agents iteratively revise their answers by referencing peer responses, aiming to converge toward a collectively optimal solution. However, Xiong et al. (2023) shows that discussions are frequently terminated before consensus is reached. In some extreme cases, we observe that agents excessively reference the answers of each other, leading to an endless loop of response swapping, ultimately preventing convergence altogether. Within competition debate, agents engage in persistent opposition by defending their own answers, and the ultimate answer is selected from the set of competing responses. While this approach may be somewhat effective for binary classification tasks, it becomes problematic in multi-class or generative tasks. In scenarios where all agents start with incorrect answers, the rigid process of mutual defense and refutation may trap the discussion in a loop of invalid reasoning, ultimately forcing the selection of an incorrect final answer. Overall, regardless of whether the debate takes a convergent or competition paradigms, existing MAD approaches suffer from a fundamental, shared limitation: **fixed process design prevents agents from handling dynamically complex tasks.**

To address this issue, agents should be allowed to adaptively choose their discussion strategies according to the situation, instead of blindly following a fixed and inflexible process. Drawing inspiration from Sense-Plan-Act(Brooks 1986), a well-established framework in robotics where agents perceive environmental dynamics to plan and perform actions adaptively. We introduce Analyze–Compose–Execute (ACE), a novel framework for implementing dynamic MAD, which consists of three core modules. **(1) Analyze:** agents examine the current responses to construct auxiliary information which supports the dynamic composition of discussion actions. This module also assesses whether the existing answers are sufficient to conclude the discussion or not. **(2) Compose:** we construct the Atomic Discussion Act Library (ADAL), which comprises a set of atomic discussion acts. Agents autonomously select and compose atomic acts, along with auxiliary information, to generate the discussion action to be executed. **(3) Execute:** agents execute the composed discussion actions to generate new answers and explanations. Compared to existing MAD approaches, our method breaks the limitations imposed by fixed pipeline with a dynamic MAD framework.

We summarize our contributions as follows.

- We propose a novel dynamic MAD framework consisting of three modules: Analyze, Compose, and Execute. This framework enables agents to dynamically compose and execute discussion actions based on the analysis of current responses.
- We design and construct an ADAL consisting of a set of atomic discussion acts, and develop a mechanism for agents to dynamically compose these discussion actions based on ADAL.
- Both qualitative and quantitative evaluations on the challenging BBH benchmark validate the effectiveness and scalability of our proposed method.

Methods

Overview

We propose a novel **ACE** framework that enables agents to flexibly solve complex and diverse problems through dynamic discussions. As shown in Figure 2. Within this framework, agents dynamically compose the discussion actions based on the current outcomes, aiming to collaboratively produce both accurate answers and coherent explanations. To facilitate this process, we divide agents into debaters and observers based on their responsibilities, denoted as $\{A_d, A_o\}$.

- **Debaters** A_d, as the agents responsible for conducting the discussion, are tasked with composing the next-round discussion actions based on the responses generated by agents during the current discussion. A_d then execute these actions to generate new answers, repeating this process until the discussion concludes.
- **Observer** A_o, in contrast, do not directly participate in the discussion, ensuring the objectivity of the generated content. Instead, A_o analyze the answers produced by A_d to identify both their similarities and differences, which serve as auxiliary information for A_d to dynamically generate discussion actions. Additionally, A_o determine whether the current answers satisfy the task requirements, thereby flexibly controlling whether the discussion should be terminated.

Specifically, our framework is composed of three modules: Analyze, Compose and Execute. Through the coordinated operation of the three modules, our framework enables dynamic MAD and flexible control over the discussion process.

Analyze Module

The correlation among the responses of A_d constitutes a key determinant in composing discussion actions. Strong correlation tends to promote the use of cooperative strategies, whereas weak correlation motivates more confrontational approaches. The Analyze module is designed to extract correlation signals from the responses of A_d , which are operationalized as answer agreement and explanation similarity.

Consequently, we require A_d to generate both answers a and explanations E given the input question q. Specially, the explanations E are structured as itemized entries $e^{(k)}$ to enable fine-grained comparison.

$$(a_i, E_i) = A_{di}(q) \qquad i \in \{1, 2, \ldots, N\}$$

$$E_i = \{e_i^{(1)}, e_i^{(2)}, \ldots, e_i^{(k)}\}$$

where N denotes the number of A_d.

Then, the observer A_o should compare the answers and explanations (a, E) provided by A_d. Specifically, we require A_o to evaluate pairs of explanation entries $e^{(k)}$ generated by different agents and produce two sets containing similar explanation entries S and dissimilar explanation entries D.

$$S^{i,j} = \{(k, g) \mid \text{sim}(e_i^{(k)}, e_j^{(g)}) = 1, i \neq j\}$$

$$D^{i,j} = \{(k, g) \mid \text{sim}(e_i^{(k)}, e_j^{(g)}) = 0, i \neq j\}$$

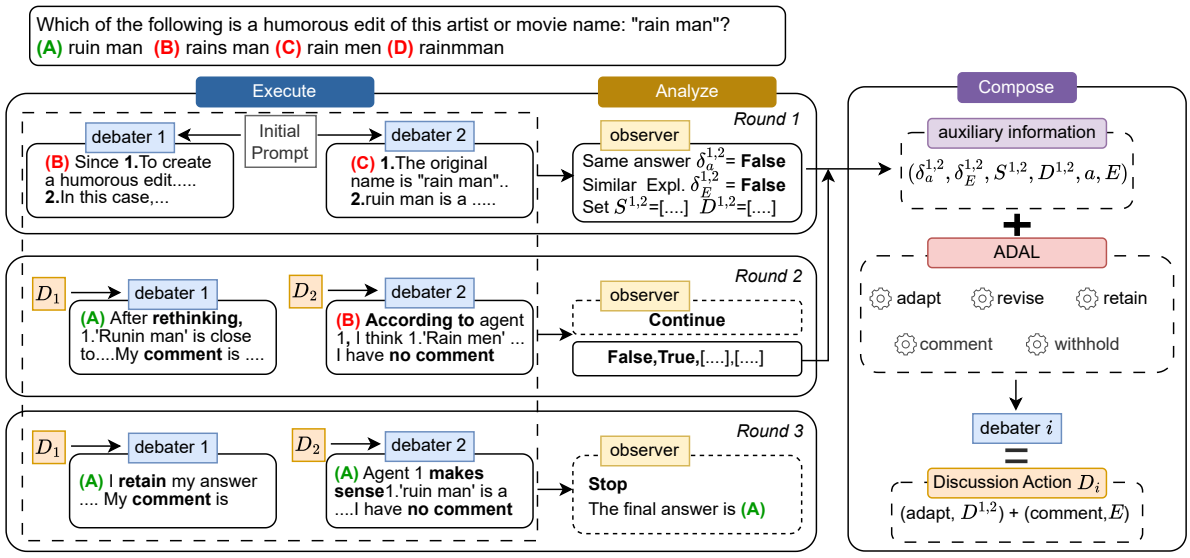

Figure 2: An overview of the ACE framework with three modules running in three discussion rounds. In the first round, the Execute module initializes the discussion by prompting debaters to directly generate their answers. The observer produces two types of outputs: solid boxes represent the observer's analysis of debater responses, while dashed boxes indicate its judgment on whether further discussion is required. Green bold text denotes correct answers, and red bold text denotes incorrect ones.

where $\text{sim}(\cdot,\cdot)$ is a function that indicates whether two explanation entries are semantically similar accomplished by A_o. Based on the sets S and D, A_o ultimately determine whether the explanations of A_d exhibit similarity $\delta_E^{(i,j)}$, and determine answer agreement $\delta_a^{(i,j)}$ by directly comparing the answers.

$$(\delta_a^{(i,j)}, \delta_E^{(i,j)}, S, D) = A_o(q || E_i || E_j)$$

$$(\delta_a^{(i,j)}, \delta_E^{(i,j)}) \in \{True, False\}$$

Additionally, A_d is responsible for analyzing whether the current responses of A_d sufficiently fulfill the task requirements, thereby determining whether further discussion is necessary.

Compose Module

We argue that achieving truly dynamic discussion requires agents to autonomously determine which discussion actions to take. However, due to the *hallucination* issues inherent in LLMs(Rawte et al. 2023), directly prompting agents to generate discussion actions may lead to highly uncontrollable and unreliable discussion trajectories. To enable controllable and dynamic discussions, we design a compositional scheme where agents construct discussion actions by selecting atomic acts from a predefined ADAL, guided by structured auxiliary information.

To construct ADAL, we decompose existing patterns of human and agent discussions and identify two fundamental dimensions of discussion acts: *Self-Oriented Acts*, whether to revise one's own answer and *Peer-Oriented Acts*, whether to provide feedback on others' answers. There are three atomic acts under self-oriented acts. **Adapt.** The agent updates its response based on the responses provided by other agents. **Revise.** The agent independently updates its answer without relying on external input. **Retain.** The agent keeps its original answer unchanged. And two atomic acts under peer-oriented acts. **Comment.** The agent provides feedback or suggestions on other agents' responses. **Withhold.** The agent does not comment on other agents' responses. To help LLMs better understand acts and reduce the gap between humans and LLMs, we provide necessary descriptions and illustrative examples for each atomic act.

Concretely, following the Analyze module, each debater A_{di} receives a set of auxiliary information $\mathcal{I}_i$ aggregated from all other debaters A_{dj} $(j \neq i)$, which includes the correlation among agents and detailed explanation matches.

$$\mathcal{I}_i = \left\{ \left(\delta_a^{(i,j)},\ \delta_e^{(i,j)},\ S^{(i,j)},\ D^{(i,j)},\ a_j,\ E_j \right) \;\middle|\; j \in \mathcal{J}_i \right\},$$

$$\mathcal{J}_i = \{ j \mid j \in [1, N],\ j \neq i \}$$

Subsequently, A_d are required to select acts from ADAL based on the $\mathcal{I}_i$ and compose them into discussion actions.

$$action = A_d(\text{ADAL} || \mathcal{I}_i)$$

For instance, when both $(\delta_a^{(i,j)},\ \delta_e^{(i,j)})$ are true, A_d determine that their responses is highly aligned with those of other agents. In this case, A_d may choose the adapt and withhold acts, using information from $D^{(i,j)}$ for the adapt act. The resulting discussion action would involve updating parts of their own responses that differ from the other agents', while withholding comments on the peers' responses. In contrast, when when both $(\delta_a^{(i,j)},\ \delta_e^{(i,j)})$ are false, indicating a strong conflict between responses, A_d may compose more assertive actions, such as rethinking its answer independently without referencing others (revise), or explicitly expressing disagreement with peer responses (comment + a_j + E_j).

Execute Module

After composing the discussion actions, A_d execute them and generate responses, which consist of the answer a, the explanation E, and comments c on the responses of other agents. To extract the final answer from our framework for performance evaluation, we prompt the A_d to output their responses in JSON format.

$$(a, E, c) = A_d(p_{\text{exe}} || \, action)$$

Prompt: You will be given an **action** describing what to do. Please execute it to generate a response with three parts:1.answer 2.explanation 3.comments(if any) Output the result in JSON format:"answer": " " ,"explanation":" " ,"comments":" " } Here is the action:{}

Subsequently, agents start a new round of discussion and iteratively follows the Analyze-Compose-Execute loop until A_o determine that the discussion can be terminated, at which point the final answer is produced.

Experiments

Tasks

The primary motivation for proposing MAD methods is to tackle complex and challenging problems which are difficult for a single LLMs. However, prior MAD approaches are often evaluated on general-purpose datasets originally designed to test the performance of individual models, which limits their ability to reflect the potential of multi-agent frameworks. Therefore, we adopt BIG-Bench Hard (BBH), a benchmark featuring more diverse and difficult problems, to better assess the capabilities of MAD methods. Additionally, to ensure fair comparison and maintain continuity with prior work, we incorporate selected subsets from datasets previously used in MAD studies as supplementary datasets alongside BBH.

BIG-Bench Hard. Suzgun et al. (2023) select a subset of 23 particularly challenging tasks on BIG-Bench and group the subset into a new benchmark referred as BIG-Bench Hard (BBH), including mathematical reasoning tasks, commonsense Understanding tasks, and scenario-based question answering. Experimental results from (Suzgun et al. 2023) indicate that a single LLM struggles to achieve satisfactory performance on BBH without the support of additional reasoning mechanisms.

Supplementary datasets The selection criteria of supplementary datasets requires that the dataset is not only representative but also serves as a complementary resource to BBH. **MMLU**, used in (Du et al. 2023), is a multi-task dataset that contains a large number of factual knowledge questions, making it a suitable complement to BBH's coverage of general knowledge. **StrategyQA**, used in (Chen, Saha, and Bansal 2023), primarily focuses on logical reasoning tasks, which are relatively underrepresented in BBH. **Commonsense Machine Translation (Common MT)**, used in (Liang et al. 2023), is a machine translation dataset featuring ambiguous translation problems that are particularly challenging for single LLMs, serves as a complement to BBH in the cross-lingual setting, addressing the absence of translation tasks in the original BBH benchmark.

Implementation Details

We follow the experimental setup of previous methods by using the zero-shot setting and avoiding the use of prompt engineering techniques such as *chain-of-thought* (CoT). To ensure consistency with baselines, which all adopt 3 agents settings, our method also employs 3 agents, including 2 debaters and 1 observer. Since all baselines require a predefined number of discussion rounds, we set the number of rounds to 4, balancing performance and cost. Additionally, to evaluate each method fairly on the BBH benchmark, we design a set of prompts that constrain the output format of LLMs. These prompts are tailored to the specific task types in BBH, such as classification, multiple choice, and text generation, to ensure that the generated answers adhere to the required formats. We conduct experiments mainly on the GPT-3.5-turbo model and set the temperature of LLMs to 0.0 following prior work for reproducibility. We did not modify other parameters, such as top_p, to ensure that the agents can still generate diverse content. All parameter settings are kept identical to previous work to ensure fairness (Du et al. 2023). All baseline methods are evaluated using their original prompts, with the only modification being the use of our format-constraining prompts to ensure consistent answer formatting on BBH. This setup is intended to faithfully reproduce their performance under fair and comparable conditions.

Baselines

To ensure the generality of our experimental results, we select both convergent and competition MAD methods as baselines.

Peer-Aware Revision (PAR)(Du et al. 2023). The PAR framework requires multiple agents to first independently answer a question, generating their initial responses. Then, a collaborative discussion phase begins, during which each agent revises its own answer by referencing the responses of others. This process reflects mutual assistance and iterative refinement among agents. The discussion continues until a predefined number of rounds is reached.

Role-Playing Debate (RPD)(Chan et al. 2023). RPD also adopts a cooperative multi-agent approach, with its main innovation being the assignment of distinct roles, such scientist, to participating agents. Notably, in RPD, these roles are intended to induce diverse perspectives, thereby enriching the debate through varied content generation. This design fundamentally differs from our approach, where roles are functionally defined.

Round Table Conference (RTC)(Chen, Saha, and Bansal 2023). Unlike other methods, RTC leverages multiple large language models of different architectures, such as ChatGPT, Bard, and Claude. These heterogeneous agents collaborate through discussion, and the final answer is com-

	BIG -Bench Hard Task	Cooperative			Competitive	Analyze–Compose–Execute (ours)
		PAR	RPD	RTC	Triadic Adversarial Reasoing	
Mathematical	Boolean Expressions	74.3±1.1	78.9±0.9	81.2±3.1	80.4±1.5	**84.3**±0.9
	Dyck Languages	25.1±0.9	24.3±1.1	**26.5**±1.5	23.1±0.9	25.9±1.1
	Multi-Step Arithmetic [Two]	33.1±1.1	35.1±0.5	36.2±1.2	32.1±0.8	**65.9**±1.0
	Navigate	54.5±0.5	56.5±0.8	60.1±1.2	58.9±2.1	**67.8**±0.4
	Object Counting	45.3±1.1	43.4±0.7	46.7±0.9	50.1±1.2	**53.8**±0.7
	Word Sorting	66.8±0.7	65.3±0.2	67.0±1.4	66.7±1.2	**71.2**±1.1
	Avg	48.2±3.9	49.6±3.4	53.1±3.0	52.0±3.1	**61.7**±2.9
Commonsense	Causal Judgement	55.9±0.7	54.8±1.1	57.9±2.1	58.1±1.7	**59.1**±1.2
	Date Understanding	50.1±2.1	52.1±1.7	53.2±1.9	49.3±0.5	**71.0**±0.3
	Disambiguation QA	**68.5**±1.9	67.1±1.1	64.5±1.8	64.2±1.5	66.3±0.5
	Formal Fallacies	48.6±1.1	50.9±0.7	**52.1**±2.0	47.1±2.0	49.2±0.3
	Geometric Shapes	27.8±1.0	25.6±1.1	28.1±1.2	26.3±0.9	**32.9**±0.7
	Hyperbaton	76.8±0.3	75.1±1.0	74.3±1.3	**77.6**±0.7	77.0±0.9
	Movie Recommendation	65.6±1.2	64.3±1.2	**67.0**±1.2	66.8±2.1	66.0±1.0
	Salient Translation Error Detection	44.8±0.5	46.1±1.1	45.2±1.1	43.9±1.9	**49.7**±1.1
	Snarks	72.9±0.2	74.8±1.1	75.1±1.2	71.4±1.8	**76.3**±0.9
	Sports Understanding	69.3±0.5	68.1±1.2	70.1±0.5	70.4±1.2	**79.0**±1.1
	Avg	58.6±3.3	57.2±2.1	57.8±3.6	56.4±1.9	**63.0**±3.1
Scenario-based	Logical Deduction (*avg*)	36.5±1.1	37.8±1.5	39.1±0.9	41.2±1.1	**54.1**±0.5
	Penguins in a Table	57.8±0.5	53.4±0.2	58.9±1.1	60.1±1.2	**70.1**±2.5
	Reasoning about Colored Objects	50.1±1.1	51.2±1.3	54.3±1.2	52.5±0.7	**74.5**±1.5
	Ruin Names	64.5±1.2	63.4±0.5	67.0±1.1	**68.9**±1.2	68.1±1.1
	Temporal Sequences	54.4±0.5	53.4±0.5	56.3±1.1	57.2±0.5	**59.1**±0.7
	Tracking Shuffled Objects (*avg*)	27.3±1.1	21.2±0.7	20.1±0.9	22.3±0.5	**37.1**±0.9
	Web of Lies	43.1±1.2	42.9±0.5	45.1±1.2	41.2±0.2	**57.1**±1.2
	Avg	47.2±3.2	46.0±3.2	48.2±3.2	49.3±2.7	**61.0**±3.7
	All Tasks (*avg*)	50.2±2.9	51.6±3.1	52.5±1.5	52.1±2.6	**61.0**±2.6

Table 1: Zero-shot prompting performance of several MAD methods on BBH. We report the mean and standard deviation performance of $Accuracy$ (%) on 23 tasks in three categories. **Best** numbers are highlighted in each column.

puted by aggregating their individual responses using a confidence-weighted scheme.

Triadic Adversarial Reasoning (TAR)(Liang et al. 2023). TAR is the only adversarial approach that adopts an competition discussion paradigm. The process begins with agents proposing their initial answers, followed by multiple rounds of mutual rebuttal. The final answer is then selected by a judge agent.

Main Results

We conduct extensive experiments to assess the performance of PAR, RPD, RTC, TAR and ACE (ours) on BBH benchmark. Overall, our method obtains the best accuracy on 17/23 tasks and shows significant improvement on the unweighted average of 23 tasks (8.5% ↑). Next, we discuss the performance presented in Table 1 divided into three categories.

Mathematical Reasoning. Some tasks in BBH, such as *Multi-Step Arithmetic* and *Obejct Counting*, require complex mathematical reasoning. *Multi-Step Arithmetic*, in particular, remains highly challenging for LLMs, which contributes to the limited performance of existing MAD approaches on such tasks. Compared to the strongest baseline, our method yields a substantial improvement of 29.7%. We attribute this to our method decomposing complex problems, enabling agents to collaborate on tasks that individual agents cannot accomplish independently. We attribute this improvement to the Analyze module in our method, which performs fine-grained analysis of responses. This facilitates mutual understanding among agents by revealing the problem-solving steps of each other , enabling them to align on correct reasoning paths while preserving diverse perspectives.

Commonsense Understanding. Some tasks in BBH require agents to possess extensive world knowledge (e.g., Movie Recommendation). We observe that on such tasks, it is often difficult for MAD methods to achieve significant gains, as LLMs already possess strong world knowledge. As a result, simply testing for commonsense recall may have limited value. However, certain tasks in BBH, such as *Date Understanding*, require LLMs to perform reasoning grounded in commonsense, rather than mere retrieval. Our method achieves a 7% improvement on *Date Understanding* and an 8% gain on *Sports Understanding*. Notably, TAR performs the worst on these tasks, likely due to the fact that they are mostly multiple-choice or generation, which poses a challenge for TAR as discussed earlier.

Scenario-based Question Answering. Several tasks in BBH simulate real-world scenarios and pose questions (e.g.,

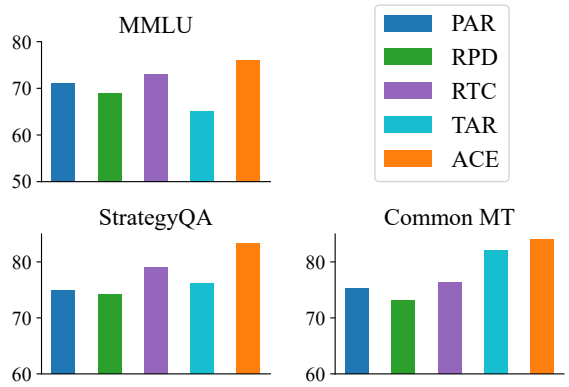

Figure 3: The performance of several MAD methods on supplementary datasets. The y-axis represents the performance of each method.

Ruin Names). These tasks are more grounded in real-world contexts, posing additional challenges for LLMs. Our method achieves the best performance on these challenging, real-world-aligned tasks, outperforming the baseline by an average of 11%. We attribute this performance to the dynamic nature of our debate framework, which allows agents to effectively adapt to diverse and evolving real-world scenarios.

Results on supplementary datasets. We conduct comprehensive experiments on the supplementary datasets, as shown in Figure 3. The evaluation metrics are kept consistent with those used in the original papers: accuracy is adopted for MMLU and StrategyQA, while COMET is used as the dynamic metric for Common MT. Notably, PAR employs MMLU to validate its method, whereas RTC does not; however, experimental results show that RTC outperforms PAR on MMLU, further supporting the potential of these supplementary datasets as effective benchmarks for evaluating MAD methods. Importantly, our ACE method achieves the best performance across all three datasets.

Ablation Study

We conducted ablation experiments on *Tracking Shuffled Objects (TSO)*, *Date Understanding (DU)*, *Penguins in a Table (PIA)* and *Geometric Shapes (GS)* tasks presented in Table 2.

***w/o* Analyze.** When the Analyze module is removed, the framework skips the analysis of the current response and directly compose discussion actions. As a result, agents lack both reference information for acts selection and the ability to operate on fine-grained discussion content. Experimental results on the PIA task show a 10% performance drop without the Analyze module, demonstrating its necessity.

***w/o* Compose.** When the Compose module is removed, the framework no longer considers how to compose actions. Instead, it randomly selects and combines atomic acts from ADAL. This leads to the most significant performance drop among all ablations, highlighting the critical role of enabling agents to reason about action composition.

Methods	Tasks			
	TSO	DU	PIA	GS
w/o Analyze	23.1±1.8	53.3±1.0	60.3±1.0	22.3±0.3
w/o Compose	21.6±0.7	47.0±0.5	49.7±1.2	17.3±1.4
w/o Execute	25.8±1.1	67.4±0.5	66.1±0.7	27.7±0.3
A + C + E	**37.1**±0.9	**71.0**±0.3	**70.1**±2.5	**32.9**±0.7

Table 2: The Ablation study on the Analyze, Compose, Execute modules.

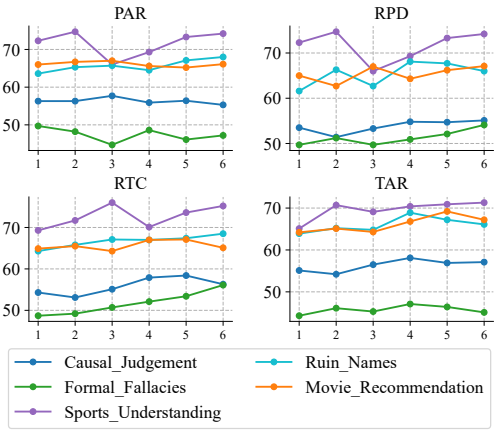

Figure 4: The impact of discussion rounds on performance. The x-axis represents the number of discussion rounds, and the y-axis represents accuracy.

***w/o* Execute.** When the Execute module is removed, the framework generates answers directly during the action composition phase, without explicitly executing the composed actions. Experimental results show that this ablation has the least impact on overall performance, though a performance drop is still observed. We attribute this to the lack of separation between action composition and execution, which increases the cognitive load on the LLMs and leads to longer context generation, thereby affecting performance.

Impact of Rounds on Performance

We evaluate the impact of the discussion rounds hyperparameter on the performance of baseline methods on *Causal Judgement (CJ)*, *Formal Fallacies (FF)*, *Sports Understanding (SU)*, *Ruin Names (RN)* and *Movie Recommendation (MR)*, see Figure 4. First, while the PAR method demonstrates optimal performance at four discussion rounds on its own dataset, as reported in the original paper, this conclusion does not hold when the method is transferred to the more complex and diverse BBH benchmark. Besides, although some results reveal patterns where a specific number of rounds leads to locally optimal performance (e.g., TAR on CJ), many others highlight the instability of this hyperparameter (e.g., RTC on SU, PAR on FF, RPD on MR). These

Model	Method	Tasks			
		TSO	DU	PIA	GS
GPT (3.5-turbo)	**ours**	**37.1**±0.9	**71.0**±0.3	**70.1**±2.5	**32.9**±0.7
GPT (4.0)	PAR	35.1±0.5	72.1±0.9	69.1±0.5	39.1±0.1
	RPD	46.2±0.1	74.0±1.1	70.2±0.3	38.2±0.3
	RTC	49.0±0.3	73.2±1.4	74.3±0.7	40.0±0.3
	TAR	42.0±0.3	67.2±1.4	76.3±0.7	39.0±0.3
	ours	**54.2**±0.5	**78.0**±0.9	**81.0**±0.8	**48.2**±0.7
LLama (2-13B)	PAR	24.1±0.7	51.2±0.5	49.9±0.2	24.5±0.3
	RPD	22.3±0.9	55.3±0.1	52.1±0.8	25.0±1.1
	RTC	23.1±0.7	57.2±0.9	58.4±1.1	26.1±0.9
	TAR	26.1±0.4	53.2±0.5	55.4±1.2	23.1±1.9
	ours	**35.9**±1.1	**66.4**±1.1	**69.5**±0.9	**36.2**±1.3

Table 3: The performance of MAD methods adopted to different LLMs (GPT-3.5-turbo, GPT-4.0 and LLama-2-13B).

findings indicate that the number of discussion rounds can significantly affect the practical performance of MAD methods. This further validates the advantage of our design, in which agents autonomously determine the discussion process without relying on a fixed rounds parameter, thereby enhancing the generalizability of our approach.

Adaptation to Alternative LLMs

Results on other LLMs are shown in the Table 3. Our method achieves the best overall performance when using GPT-4.0. Notably, a comparison between our method on GPT-3.5-turbo and the baseline on GPT-4.0 reveals that our approach matches or surpasses the baseline in some cases,such as on task TSO, our method performs comparably to PAR, demonstrating its effectiveness. Furthermore, strong performance on the open-source model LLama-2-13B indicates the generality of our method across different model backbones.

Related Work

Large Language Models

Large Language Models (LLMs), often containing hundreds of billions of parameters, have become a promising path toward General Artificial Intelligence (GAI) (Wang et al. 2023b; Agrawal et al. 2022). Modern LLMs typically consist of an enormous number of parameters, such as GPT-3(Brown et al. 2020b), GLM(Zeng et al. 2023),LLaMA(Touvron et al. 2023) and Galactica(Taylor et al. 2022). Through the expansion of model size and training data, LLMs have demonstrated outstanding capabilities on downstream tasks, in some cases matching or outperforming humans. To further unleash the LLMs' potential, prompt engineering(Liu et al. 2023) is extensively employed, which significantly enhances model performance without altering the model parameters themselves, thus markedly reducing the costs of fine-tuning. Moreover, to further enhance model performance on complex tasks, prompts constructed via the Chain of Thought (COT)(Wei et al. 2022) approach are utilized to guide the LLMs in accomplishing intricate reasoning tasks. However, these methods are designed specifically for individual LLMs.

Multi-Agent Debate

Inspired by "The Society of Mind" (Minsky 1988), Du et al. (2023) has proposed a novel debate approach for large models , where a single large model iteratively refines its reasoning process by reading the outputs of other models multiple times, ultimately converging on a more accurate answer. Similarly, Tencent has introduced a Multi-Agent Debate (MAD) framework (Liang et al. 2023), wherein multiple agents articulate their arguments while a judge orchestrates the debate to reach a conclusive resolution. This framework addresses the Degeneration-of-Thought (DoT) problem that can occur when a single model undergoes self-iteration. Furthermore, Du et al. (2023) has advanced the field by utilizing an enhanced language response methodology wherein multiple language model instances engage in iterative debate to refine their contributions, culminating in a consensus final answer. In addition, Chan et al. (2023) has constructed ChatEval, a multi-agent referee team, to autonomously deliberate and assess the quality of responses from various models on open-ended questions and Natural Language Generation (NLG) tasks. Recently, Chen, Saha, and Bansal (2023) has proposed ReConcile, multi-agent framework that emulates a round table conference, enhancing collaborative reasoning among diverse LLM agents through iterative discussions, persuasive strategies, and a confidence-weighted voting mechanism to achieve a more robust consensus. To further unleash the potential of cognitive synergy in Large Language Models (LLMs), Wang et al. (2023c) has proposed Solo Performance Prompting (SPP), which dynamically assigns multiple fine-grained personas to LLMs based on task inputs. This approach collaboratively combines the strengths and knowledge of various perspectives to effectively reduce factual hallucination while maintaining the robust reasoning capabilities of the original model (Chen et al. 2024). These studies have amply demonstrated that multi-agent systems can enhance individual performance by emulating the cognitive synergy observed in human cognition, thereby addressing more complex reasoning tasks (Zhao et al. 2024).

Conclusion

This paper presents a systematic analysis of existing MAD methods and observes that they typically require agents to repeat a fixed discussion pattern until termination. Such rigid discussion workflows limit their effectiveness on complex tasks. Therefore, we propose ACE, a framework designed to enable dynamic discussions. Our method enables truly flexible and dynamic discussions by analyzing the responses of agents, selecting atomic operations and composing them into actions, then executing these actions.

In future work, we plan to further refine the ADAL library and explore methods for dynamically constructing it. Additionally, we will focus on optimizing the framework to reduce computational overhead. In future work, we plan to further refine the ADAL library and explore methods for its dynamic construction. Additionally, we aim to optimize the framework to achieve better performance.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFF0902701; in part by the National Natural Science Foundation of China under Grants U21A20468, 62372058, U22A2026; in part by the Natural Science Foundation of Anhui under Grant 2508085QF241.

References

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

Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; and Sontag, D. 2022. Large language models are few-shot clinical information extractors. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1998–2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Brooks, R. A. 1986. A robust layered control system for a mobile robot. *IEEE J. Robotics Autom.*, 2(1): 14–23.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020a. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020b. Language Models are Few-Shot Learners. In *Proceedings of NeurIPS*.

Cai, Y.; Gu, Z.; Du, Z.; Ye, Z.; Cao, S.; Xu, Y.; Feng, H.; and Chen, P. 2025. MIRAGE: Exploring How Large Language Models Perform in Complex Social Interactive Environments. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 14–40. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-252-7.

Chan, C.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *arXiv preprint arXiv:2308.07201*.

Chawla, K.; Wu, I.; Rong, Y.; Lucas, G.; and Gratch, J. 2023. Be Selfish, But Wisely: Investigating the Impact of Agent Personality in Mixed-Motive Human-Agent Interactions. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13078–13092. Singapore: Association for Computational Linguistics.

Chen, G.; Dong, S.; Shu, Y.; Zhang, G.; Sesay, J.; Karlsson, B. F.; Fu, J.; and Shi, Y. 2024. AutoAgents: A Framework for Automatic Agent Generation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 22–30. Main Track.

Chen, J. C.; Saha, S.; and Bansal, M. 2023. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. *arXiv preprint arXiv:2309.13007*.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Qian, C.; Chan, C.-M.; Qin, Y.; Lu, Y.; Xie, R.; et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; and Agrawal, S. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325*.

Fan, C.; Chen, J.; Jin, Y.; and He, H. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press. ISBN 978-1-57735-887-9.

Gu, Z.; Zhu, X.; Guo, H.; Zhang, L.; Cai, Y.; Shen, H.; Chen, J.; Ye, Z.; Dai, Y.; Gao, Y.; Hu, Y.; Feng, H.; and Xiao, Y. 2024. Agent Group Chat: An Interactive Group Chat Simulacra For Better Eliciting Collective Emergent Behavior.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; Xiao, L.; Wu, C.; and Schmidhuber, J. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*.

Hua, W.; Fan, L.; Li, L.; Mei, K.; Ji, J.; Ge, Y.; Hemphill, L.; and Zhang, Y. 2023. War and Peace (WarAgent): Large Language Model-based Multi-Agent Simulation of World Wars.

Li, H.; Chong, Y.; Stepputtis, S.; Campbell, J. P.; Hughes, D.; Lewis, C.; and Sycara, K. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 180–192.

Li, N.; Gao, C.; Li, M.; Li, Y.; and Liao, Q. 2024. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15523–15536. Bangkok, Thailand: Association for Computational Linguistics.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Tu, Z.; and Shi, S. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *arXiv preprint arXiv:2305.19118*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*.

Lu, S.; Shao, J.; Luo, B.; and Lin, T. 2025. MorphAgent: Empowering Agents through Self-Evolving Profiles and Decentralized Collaboration. arXiv:2410.15048.

Minsky, M. 1988. *Society of mind*. Simon and Schuster.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18.

Rawte, V.; Chakraborty, S.; Pathak, A.; Sarkar, A.; Islam Tonmoy, M. T., S. Chadha, A.; Sheth, A. P.; and Das, A. 2023. The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2541–2573. Singapore.

Siyu Li, J. Y.; and Zhao, K. 2023. Are you in a Masquerade? Exploring the Behavior and Impact of Large Language Model Driven Social Bots in Online Social Networks. *arXiv preprint arXiv:2307.10337*.

Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Ouyang, X.; Yu, D.; Tian, H.; Wu, H.; and Wang, H. 2021. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *arXiv preprint arXiv:2107.02137*.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; and Wei, J. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Proceedings of ACL*.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A Large Language Model for Science. *arXiv preprint arXiv:2211.09085*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2309.16609*.

Touvron, H.; and Martin, L. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Tran, K.; Dao, D.; Nguyen, M.; Pham, Q.; O'Sullivan, B.; and Nguyen, H. D. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *CoRR*, abs/2501.06322.

Wang, K.; Lu, Y.; Santacroce, M.; Gong, Y.; Zhang, C.; and Shen, Y. 2025. Adapting LLM Agents with Universal Communication Feedback. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 6090–6107. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.

Wang, L.; Lyu, C.; Ji, T.; Zhang, Z.; Yu, D.; Shi, S.; and Tu, Z. 2023a. Document-Level Machine Translation with Large Language Models. In *Proceedings of NeurIPS*.

Wang, Y.; Le, H.; Gotmare, A.; Bui, N.; Li, J.; and Hoi, S. 2023b. CodeT5+: Open Code Large Language Models for Code Understanding and Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1069–1088. Singapore: Association for Computational Linguistics.

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

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of NeurIPS*.

Xiong, K.; Ding, X.; Cao, Y.; Liu, T.; and Qin, B. 2023. Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate. In *Proceedings of EMNLP*.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; Tam, W. L.; Ma, Z.; Xue, Y.; Zhai, J.; Chen, W.; Liu, Z.; Zhang, P.; Dong, Y.; and Tang, J. 2023. GLM-130B: An Open Bilingual Pre-trained Model. In *Proceedings of ICLR*.

Zhang, C.; Yang, K.; Hu, S.; Wang, Z.; Li, G.; Sun, Y.; Zhang, C.; Zhang, Z.; Liu, A.; Zhu, S.-C.; Chang, X.; Zhang, J.; Yin, F.; Liang, Y.; and Yang, Y. 2024a. ProAgent: Building Proactive Cooperative Agents with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17591–17599.

Zhang, J.; Xu, X.; Zhang, N.; Liu, R.; Hooi, B.; and Deng, S. 2024b. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14544–14607. Bangkok, Thailand: Association for Computational Linguistics.

Zhao, Q.; Wang, J.; Zhang, Y.; Jin, Y.; Zhu, K.; Chen, H.; and Xie, X. 2024. CompeteAI: understanding the competition dynamics of large language model-based agents. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.