Explain-Analyze-Generate: A Sequential Multi-Agent Collaboration Method for Complex Reasoning

Wenyuan Gu[1*], **Jiale Han**[2*], **Haowen Wang**[3], **Xiang Li**[1] **and Bo Cheng**[1†]
[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2]Hong Kong University of Science and Technology
[3]School of Computer Science and Technology, Anhui University
{guwenyuan,lixiang2022,chengbo}@bupt.edu.cn,jialehan@ust.hk,wanghaowen@ahu.edu.cn

Abstract

Exploring effective collaboration among multiple large language models (LLMs) represents an active research direction, with multi-agent debate (MAD) emerging as a popular approach. MAD involves LLMs independently generating responses and refining their own responses by incorporating feedback from other agents in a debate manner. However, empirical experiments reveal the suboptimal performance of MAD in complex reasoning scenarios. We attribute this to the potential misleading caused by peer agents with limited individual capabilities. To address this, we propose a novel sequential collaboration framework named Explain-Analyze-Generate (EAG). By decomposing complex tasks into essential subtasks and employing a pipeline approach, EAG enable agents provide constructive assistance to peers, ultimately yielding higher performance. We conduct experiments on the comprehensive complex language reasoning benchmark: BIG-Bench-Hard (BBH). Our method achieves the highest performance on 19 out of 23 tasks, with an average improvement of 8% across all tasks, and incurs lower costs compared to MAD, demonstrating its effectiveness and efficiency.

1 Introduction

Large language models (LLMs) (Brown et al., 2020a; Touvron et al., 2023a,b; Achiam et al., 2023; Chowdhery et al., 2023) have made significant breakthroughs in the field of natural language processing, which achieve notable success in language understanding and generation. (Sun et al., 2021; Wang et al., 2023) However, the reasoning abilities of LLMs still present challenges (Zhu et al., 2023; Gou et al., 2023), which drives research towards enabling LLMs to mimic human cognitive behaviors (Wei et al., 2022; Madaan et al., 2023), aiming to prompt LLM reasoning ability through human-like task-solving strategies.

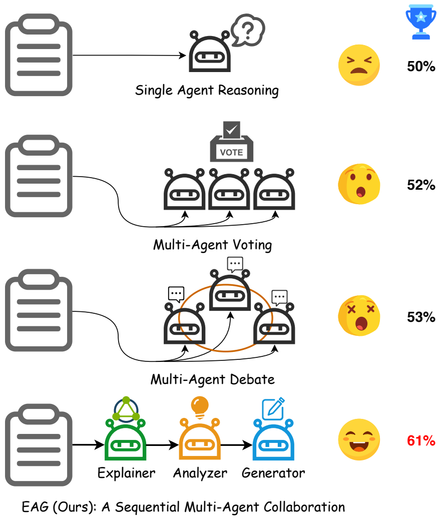


Figure 1: Results on BBH (Suzgun et al., 2023).

Inspired by the concept of society of minds (Minsky, 1988) in multi-agent systems, multi-agent debate (MAD) (Liang et al., 2023; Du et al., 2023) has been proposed. In specifical, a query is simultaneously fed into different LLMs and each LLM generates candidate answers independently. Subsequently, each LLM reviews the responses from all other LLMs to refine its own answer. This process is repeated for several rounds, culminating in a final answer. MAD has been a popular multi-agent collaboration paradigm and widely applied to sentiment analysis (Sun et al., 2023), text evaluation (Chan et al., 2023), and AI systems (Li et al., 2023). Despite notable achievements, previous works predominantly center on tasks that require minimal reasoning, which may limit MAD from reaching its full capability.

To explore the maximum potential of MAD, we adopt a challenging reasoning benchmark BIG-Bench-Hard (BBH) (Suzgun et al., 2023) to examine its performance, which encompasses 23 diverse

[*] These authors contributed equally to this work.
[†] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7127–7140
January 19–24, 2025. ©2025 Association for Computational Linguistics

tasks including logical reasoning, mathematical computation, common-sense understanding, and scenario simulation. BBH is more closely aligned with real-world complexities than general tasks and poses greater challenges for language models. As presented in Figure 1, we compare the performance of MAD with single-agent reasoning and multi-agent voting, the latter aggregating the majority answer of multiple agents as the final answer without debate. Despite the expensive and time-consuming multiple rounds of debate, surprisingly, MAD fails to deliver the significant improvements as expected compared to multi-agent voting. Through an in-depth empirical analysis of MAD (detailed in Section 2), we discover that individual LLMs have limited capacity to handle challenging problems and are prone to generating incorrect answers, which consequently mislead other LLMs during the debate process and ultimately result in unsatisfactory performance. We attribute this to the parallel collaborative nature of MAD, that is, although agents receive thoughts and potential answers from their peers as references, multiple agents are independently responsible for solving the entire task, which is highly challenging especially for complex reasoning tasks.

Taking a closer look at the limitation of single agents in handling complex reasoning tasks, we observe that the poor performance of single agents can be attributed to inaccurate question understanding, faulty reasoning steps, and inconsistent answers with reasoning. In light of the above findings, we propose Explain-Analyze-Generate (EAG), a novel sequential multi-agent collaboration method for complex reasoning, which decomposes complex tasks to simple and critical subtasks and allocates each subtask to a single agent for sequential division of labor. In specific, EAG framework comprises three core components. **(1) Explainer** focuses on key information and clarifies the true intent of the questions to help subsequent agents better understand and solve problems. **(2) Analyzer** cuts open problems and proposes reasoning insights and solution approaches based on the key information from the explainer. **(3) Generator** aggregates insights and solutions and executes responses that adhere to reasoning and specified output formats. Compared to MAD, this sequential collaborative framework mirrors human teamwork and distributes a lighter workload to individual agents, enabling them to provide more specific solutions to their individual tasks and offer positive assistance to other LLMs, ultimately yielding higher performance. We conduct extensive experiments under the zero-shot setting of BBH tasks and the results demonstrate the effectiveness and efficiency of our method. We release our code[*] to the community for future research.

We summarize our contributions as follows:

- Our comprehensive experiments on complex reasoning unveil that the bottleneck for unsatisfactory performance of MAD lies in the misleading induced by the wrong thoughts of other agents. The parallel collaborative nature of MAD overburden single LLMs, hindering its effectiveness on challenging tasks.
- We propose a novel sequential multi-agent collaboration framework EAG, which employs explainer, analyzer, and generator for cooperative division of labor, fully harnessing the potential of multiple agents to effectively address confused understanding, faulty reasoning, and inconsistent responses.
- Qualitative and quantitative experiments on challenging BIG-Bench-Hard benchmark demonstrate the effectiveness of our proposed method.

2 Analysis of Multi-Agent Debate

Given a question q and N agents $\{A_i, i = 1, 2, ...N\}$, the general multi-agent debate framework includes three stages. *1) Initial Response:* Each agent A_i independently generates its own answer a_i^0 to the question q. *2) Multi-Round Debate:* multiple agents engage in R rounds of debate. In the r-th round, each agent A_i receives the output of the other agents $\{a_j^{r-1}, j = 1, 2, ..., N, j \neq i\}$ from preceding rounds to refine and revise its own answer a_i^r. *3) Final Answer:* after R rounds of debate, MAD chooses the majority answer among the agents $\{a_i^R, i = 1, 2, ..., N\}$ as the final answer a, as multiple agents do not always converge and achieve consensus.

We adopt multi-agent voting as a baseline for comparison with MAD. In this method, several agents independently answer the question, and the final answer is determined by the majority vote. Clearly, the results of multi-agent voting are the same as the initial response of MAD. The experimental results of multi-agent voting and MAD on

[*] https://github.com/buptgwy/Explain-Analyze-Generate

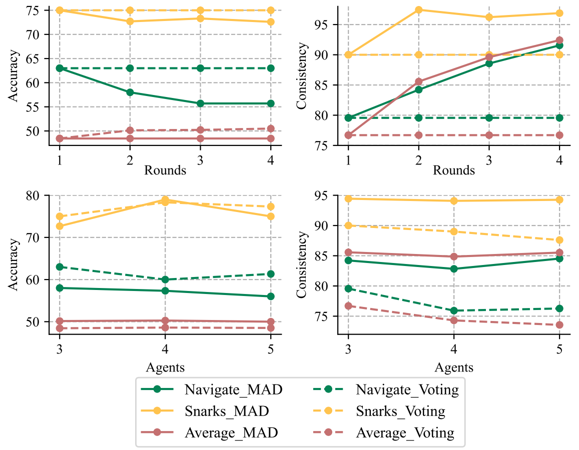
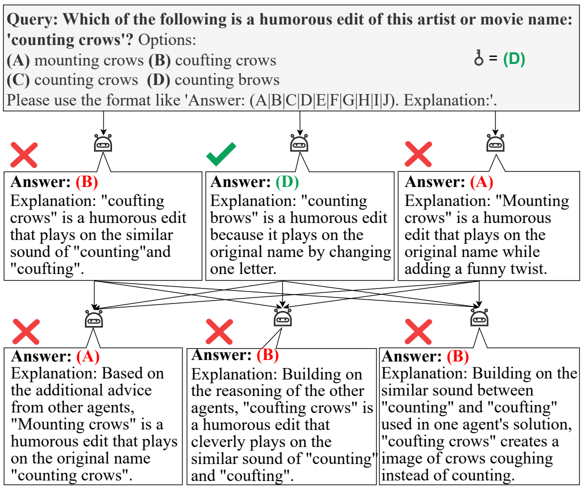


Figure 2: Accuracy and consistency changes of multi-agent voting and MAD across different rounds and agent counts on BBH. Both *Snark* and *Navigate* are tasks in BBH. *avg*: unweighted average of all tasks in BBH. See Appendix A.1 for detailed scores.


Query: Which of the following is a humorous edit of this artist or movie name: 'counting crows'? Options:
(A) mounting crows **(B)** coufting crows
(C) counting crows **(D)** counting brows
Please use the format like 'Answer: (A|B|C|D|E|F|G|H|I|J). Explanation:'.

= (D)

✗ **Answer:** (B)
Explanation: "coufting crows" is a humorous edit that plays on the similar sound of "counting"and "coufting".

✓ **Answer:** (D)
Explanation: "counting brows" is a humorous edit because it plays on the original name by changing one letter.

✗ **Answer:** (A)
Explanation: "Mounting crows" is a humorous edit that plays on the original name while adding a funny twist.

✗ **Answer:** (A)
Explanation: Based on the additional advice from other agents, "Mounting crows" is a humorous edit that plays on the original name "counting crows".

✗ **Answer:** (B)
Explanation: Building on the reasoning of the other agents, "coufting crows" is a humorous edit that cleverly plays on the similar sound of "counting" and "coufting".

✗ **Answer:** (B)
Explanation: Building on the similar sound between "counting" and "coufting" used in one agent's solution, "coufting crows" creates a image of crows coughing instead of counting.

Figure 3: Example of misleading phenomenon.

BBH are shown in Figure 2. These results show several interesting findings, which we briefly summarize here:

MAD doesn't show satisfactory improvement and consistency despite costly debates. We are surprised to find that, despite consuming more tokens and time, MAD only achieves performance comparable to multi-agent voting, and in some cases, even underperforms, particularly on tasks such as *snarks* and *navigate*. Moreover, the influence of debate rounds on MAD exhibits inconsistency, as shown in the upper part of Figure 2. As the number of debate rounds increases, the performance of MAD does not steadily improve as expected, instead displaying fluctuations, indicating that MAD lacks robustness. In addition, from the right part of Figure 2, we notice that multiple agents fail to reach a consensus after debates, which suggests that the confidence of debate outcomes remains uncertain and the full engagement of each agent within the debate framework is not ensured.

The misleading induced by the wrong thoughts of other agents is to blame. In investigating the unsatisfactory performance of MAD, we examine the debate responses among agents, identifying a prevalent misleading phenomenon. We illustrate a specific case of misleading in Figure 3. Initially, *Agent 2* provides the correct answer based on individual reasoning. However, in subsequent debate rounds, *Agent 2* is misled by *Agent 3* and adopts the same incorrect answer. We further conduct a statistical analysis on the frequency of the misleading phenomenon across all BBH tasks. First we count the number of agents providing correct answers, and then calculate the proportion of these agents who subsequently provided incorrect answers in the following debate rounds. The frequency of the misleading phenomenon under varying agent number and debate round settings is presented in the Table 1. It can be observed that the frequency of misleading is notably high, consistently exceeding 18% across different settings, and the misleading phenomenon accumulates with the increase of debate rounds.

Let's take a closer look at the potential causes of the misleading. For agents whose answers reverse from true to false (T2F), we compute the proportion of responses from the other agents that are incorrect. The results are summarized in Table 2. Notably, the proportions in the T2F scenario exceed 60%, indicating the answer misleading of a single agent during the debates primarily comes from the incorrect thoughts from the other agents. To further validate this conclusion, we conduct a converse experiment to examine agents with a reversal from false to true (F2T) and calculate the proportion of correct responses from other agents. From Table 2, we observe the same phenomenon as the T2F scenario, that is, more than half of the opinions from other agents are correct, which strongly supports our conclusion that the agent's answers during MAD discussion are biased by the majority opinions of the other agents. Based on the above analysis, we come to a hypothesis: *In complex reasoning tasks, a single agent is insufficient to solve the problem independently and prone to generat-*

Table 1: Proportion of instances that were correct in the first round but incorrect in subsequent rounds, relative to the total number of correct instances initially observed in the first round.

Agents \Rounds	2	3	4
3	19.3%	29.2%	30.6%
4	18.5%	-	-
5	18.5%	-	-

Table 2: Proportion of *True to False (T2F)* and *False to True (F2T)* changes across different rounds and agent counts.

Number	Rounds=2		Rounds=3	
of Agents	*T2F*	*F2T*	*T2F*	*F2T*
3	70.1%	60.8%	73.6%	70.6%
4	67.0%	54.9%	-	-
5	62.5%	52.4%	-	-

ing wrong answers, which consequently misleads other agents during the debate process This suggests that the costly debate paradigm may not be a good solution for complex reasoning tasks.

The poor performance of single agents can be attributed to inaccurate understanding, faulty reasoning, and inconsistent answers. We randomly sample 200 instances from BBH with incorrect predictions to investigate the reasons behind agents' subpar performance in MAD. After manual analysis of errors by three volunteers with well-trained backgrounds, we categorize the reasons for erroneous answers into three groups. *1) Inaccurate question understanding.* The information redundancy of queries pose challenges for single agents in accurately discerning the question's true intent. *2) Faulty reasoning steps.* Single agent produces incomplete or incorrect reasoning steps due to their limited capacity to address the entire task. *3) Inconsistent answers with reasoning.* Agents reason step-by-step correctly but provide wrong final answers. The proportions of the three error types are presented in Figure 4. Detailed examples of the three error types are listed in Appendix A.2. This discovery inspires us to design a sequential multi-agent collaboration method, utilizing aggregation, analysis, and answer agents to collaborate in a division of labor, mitigating the limitations of individual agents.

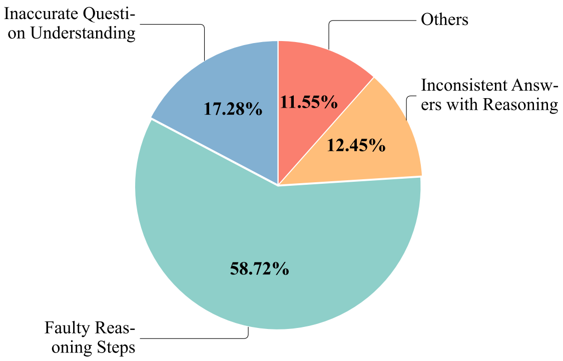


Figure 4: Proportions of Different Error Types Identified in Agent Performance Analysis. See Appendix A.2 for detailed error cases.

3 Method

Based on the above analysis, we present a sequential multi-agent collaboration framework called Explain-Analyze-Generate (EAG) for complex reasoning tasks, as depicted in Figure 5. This framework comprises three key agents, denoted as $\{A_e, A_a, A_g\}$, which collaborate sequentially. Given a question q, explainer A_e firstly reads and organizes information derived from the question, followed by analyzer A_a, who reasons and processes about the complex questions. Finally, generator A_g synthesizes the information and analysis from A_e and A_a and generates the answer. By doing so, the complex reasoning problem is broken down into three critical and easy subtasks, and is gradually solved through sequential collaboration among multiple agents. Compared to MAD, EAG only requires to engage in one round of sequential collaboration, without the need for multiple rounds of discussion, making it more efficient and stable.

Explainer Complex tasks often contain cognitive traps and noisy information, making it difficult for models to accurately understand the true intent of the problem. Therefore, focusing on key details, filtering out irrelevant information, and discerning the true purpose of complex problems are crucial for guiding task-solving approaches. Consequently, we propose an explainer A_e to clarify the problem's intent and provide guidance on solving it, which helps subsequent agents better understand the task. This aids subsequent agents in gaining a clearer understanding of the task at hand. We prompt A_e to generate its explanation y_{A_e} given

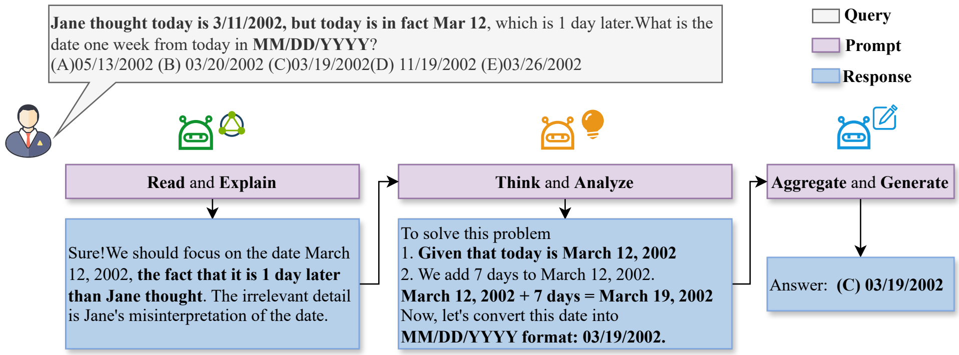

Jane thought today is 3/11/2002, but today is in fact Mar 12, which is 1 day later.What is the date one week from today in MM/DD/YYYY?
(A)05/13/2002 (B) 03/20/2002 (C)03/19/2002(D) 11/19/2002 (E)03/26/2002
Query
Prompt
Response
Read and Explain
Think and Analyze
Aggregate and Generate
Sure!We should focus on the date March 12, 2002, the fact that it is 1 day later than Jane thought. The irrelevant detail is Jane's misinterpretation of the date.
To solve this problem
1. Given that today is March 12, 2002
2. We add 7 days to March 12, 2002.
March 12, 2002 + 7 days = March 19, 2002
Now, let's convert this date into MM/DD/YYYY format: 03/19/2002.
Answer: (C) 03/19/2002

Figure 5: Overview on EAG method. 🤖 indicates explainer, 🤖 indicates analyzer and 🤖 indicates generator. We simplify the prompts of each agent for brevity. More details of prompt see Appendix A.3.

the input question q and prompt p_{A_e}:

$$y_{A_e} = A_e(p_{A_e} \,||\, q)$$

where $||$ denotes concatenation, and prompt p_{A_e} is as follows:

> **Prompt:** Please review and share your understanding of the question to assist the following agents in improving their problem-solving abilities. Ensure you focus on essential information while filtering out irrelevant details.

🤖 **Analyzer** The Analyzer agent A_a is entrusted with divergently analyzing problems and proposing solution approaches. Specifically, given q, prompt p_{A_a}, and information y_{A_e} from A_e, A_a generates thoughts for solving q. Providing A_a with core information y_{A_e} relevant to the task enables it to better understand the complex question, thereby improving its logical reasoning capabilities.

$$y_{A_a} = A_a(p_{A_a} \,||\, q \,||\, y_{A_e})$$

Prompt p_{A_a} is as follows:

> **Prompt:** Please review the question and the understanding from the previous agent, and provide a comprehensive reasoning process to solve the problem. Ensure clarity and detail in your thought process to effectively guide the following agent in generating the final answer.

🤖 **Generator** When answering questions, LLMs tend to provide all the relevant knowledge they possess, resulting in final answers that may be imprecise or not aligned with human inclination. Therefore, we design an generator agent A_g to integrate key information with the thought process and generate final answers that meet task requirements. Given q, prompt p_{A_g}, and outputs from A_e and A_a, A_g generates the final answer y to the question q.

$$y = A_g(p_{A_g} \,||\, q \,||\, y_{A_e} \,||\, y_{A_a})$$

Prompt p_{A_g} is as follows:

> **Prompt:** Please review the question and the thoughts of previous agents, and provide the final answer to the question. Ensure to follow this format: "Answer: (A|B|C|D|E|F|G|H|I|J)".

4 Experiments

4.1 Experimental Setup

Benchmark Suzgun et al. (2023) select a subset of 23 particularly challenging tasks on BIG-Bench (Srivastava et al., 2022) and group the subset into a new benchmark referred as BIG-Bench Hard (BBH) [†], including Mathematical Reasoning tasks (Math), Natural Language Understanding tasks (NLU), and Scenario-based Question Answering tasks (ScenarioQA). The statistics and other details of BBH can be found in Appendix A.4. We follow many prior works (Du et al., 2023; Bian et al., 2023; Besta et al., 2024; Yao et al., 2023; Chen et al., 2023) and experiment with a subset of 100 samples on each task in BBH.

Implementation Details We implement our method on zero-shot setting and do not introduce any other prompt engineering technologies such as *Chain of Thought* (COT) (Wei et al., 2022). To accurately parse the answers from the responses of LLMs, we design a suit of result format prompts based on different types of questions, including multiple-choice questions, true/false questions and

[†] https://github.com/suzgunmirac/BIG-Bench-Hard

Table 3: Zero-shot prompting performance of several multi-agent collaboration methods on BBH. We report the mean and standard deviation performance of $Accuracy$ (%) on 23 tasks in three categories. R: debate rounds. **Best** numbers are highlighted in each column.

	BIG-Bench Hard Task	Single-Agent	Multi-Agent Voting	Multi-Agent Debate			EAG (ours)
				R=2	R=3	R=4	
Mathematical	Boolean Expressions	72.3±1.0	79.7±1.0	80.0±2.4	81.7±1.4	79.7±1.5	**85.3**±1.2
	Dyck Languages	15.3±2.1	21.3±1.7	**26.7**±1.1	25.0±1.4	22.7±1.0	22.0±0.4
	Multi-Step Arithmetic [Two]	2.0±0.5	2.3±0.5	28.3±1.8	33.0±1.7	34.7±1.8	**71.7**±2.0
	Navigate	57.7±1.2	63.0±0.9	58.0±2.4	55.7±1.2	54.7±1.4	**69.0**±0.9
	Object Counting	44.7±1.4	47.9±0.4	45.7±1.3	46.0±1.6	47.3±0.9	**54.0**±1.6
	Word Sorting	56.3±0.2	65.3±1.5	67.3±0.5	67.0±1.6	66.0±2.3	**70.0**±1.9
	Avg	41.2±4.3	46.6±4.2	51.0±4.0	51.4±3.8	50.9±3.9	**62.0**±3.9
Commonsense	Causal Judgement	56.3±0.3	56.3±1.2	**57.7**±1.0	57.0±2.4	55.0±1.4	57.3±1.2
	Date Understanding	50.7±1.4	55.0±0.5	58.3±1.5	48.3±0.7	57.7±0.7	**72.0**±2.9
	Disambiguation QA	67.0±0.9	67.7±0.7	65.7±1.2	67.3±1.2	65.3±1.0	**68.3**±0.8
	Formal Fallacies	49.7±2.1	48.2±1.4	44.7±4.1	45.0±3.8	44.3±2.6	**53.3**±0.5
	Geometric Shapes	26.0±0.8	27.9±0.5	28.0±0.5	29.0±1.4	29.0±0.8	**31.7**±1.4
	Hyperbaton	75.3±1.0	77.3±0.5	77.6±1.0	77.7±0.3	76.5±0.5	**78.0**±0.5
	Movie Recommendation	66.0±1.4	**66.7**±0.7	67.0±0.8	64.7±0.7	65.0±0.9	66.7±2.2
	Salient Translation Error Detection	45.0±0.8	44.0±0.8	**45.3**±0.3	45.3±1.0	45.3±1.0	44.3±0.7
	Snarks	75.0±0.5	75.0±0.8	72.7±1.2	73.3±2.0	71.6±2.2	**75.7**±1.9
	Sports Understanding	72.3±0.7	74.7±1.4	66.0±1.2	72.3±1.4	72.0±1.7	**77.7**±0.5
	Avg	58.3±2.3	59.3±2.5	58.3±2.6	58.0±2.4	58.17±3.1	**62.5**±2.6
Scenario-based	Logical Deduction (*avg*)	36.2±1.9	42.8±1.1	38.4±3.2	37.7±4.4	38.3±1.7	**52.0**±1.7
	Penguins in a Table	50.7±2.0	55.0±1.7	60.7±1.4	60.0±0.5	60.7±0.3	**71.3**±4.4
	Reasoning about Colored Objects	43.7±0.7	47.3±1.0	47.3±1.8	48.7±1.5	49.0±2.0	**69.3**±4.6
	Ruin Names	63.6±0.5	65.3±0.7	65.7±1.2	66.0±0.5	65.8±1.0	**68.0**±0.9
	Temporal Sequences	44.0±1.3	45.0±2.2	47.7±1.0	47.0±1.4	48.3±0.7	**57.0**±0.8
	Tracking Shuffled Objects (*avg*)	23.6±2.2	24.3±1.3	23.3±1.6	23.3±1.4	21.9±1.7	**34.4**±2.9
	Web of Lies	53.3±0.7	51.3±2.0	50.7±0.7	45.0±0.8	49.7±1.0	**60.3**±2.9
	Avg	45.0±3.0	47.3±2.9	47.7±3.1	46.8±2.7	47.7±3.0	**58.9**±3.2
	All Tasks (*avg*)	49.9±3.9	52.3±3.9	53.1±3.5	52.9±3.6	53.0±3.6	**61.3**±3.6

text generation questions. Details about prompts can be found in Appendix A.5.

LLMs settings and Prompts We conduct experiments mainly on the GPT-3.5-turbo model (Kojima et al., 2022). We adopt the same parameters and prompt settings as Du et al. (2023) to replicate their results on BBH. We set the temperature of our method to 0.0 following prior work (Xiong et al., 2023) for reproducibility.

4.2 Baselines

Single-Agent Reasoning (SAR) a method wherein agents, based on LLMs, directly respond to queries within tasks. We prompt agents to provide responses accompanied by explanatory rationales delineating the reasoning processes in addressing the questions.

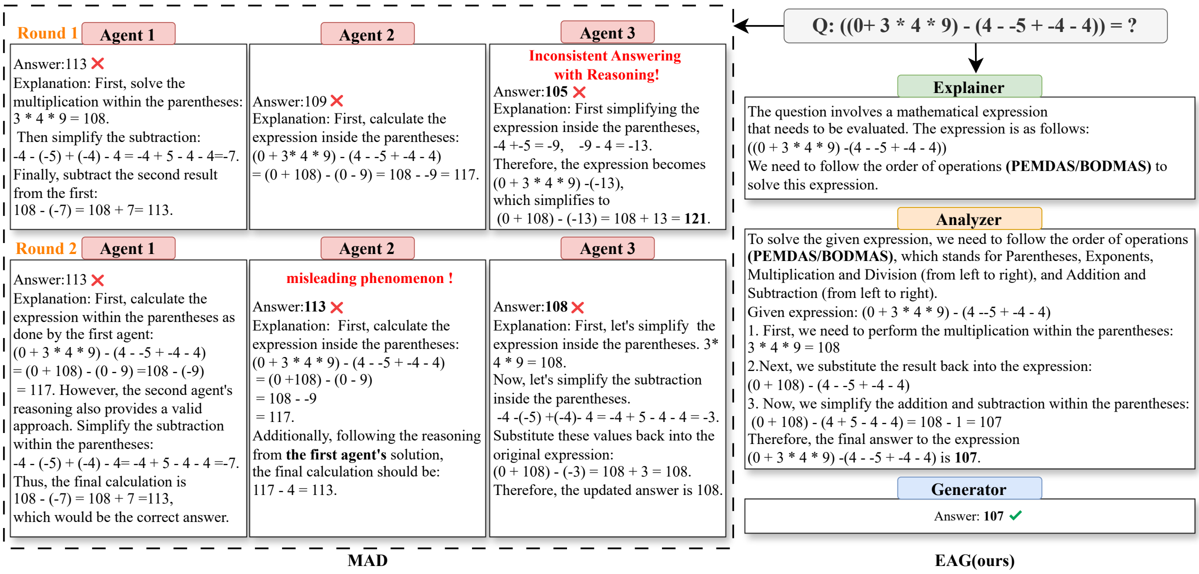
Multi-Agent Voting (MAV) a method wherein multiple agents generate response as described in single agent reasoning method at the same time, then the majority answer among the agents chosen as the final answer. Multi-agent voting is similar to the Collaboration-Soft method proposed by Xiong et al. (2023).

Multi-Agent Debate (MAD) (Du et al., 2023) a collaboration method wherein multi-agents engage in continuous debate to generate a more reasonable response. More details about MAD are described in Section 2. Moreover, we conduct experiments within MAD across varying rounds R, including R=3, 4, and 5 presented in Table 3 for a more exhaustive analysis.

4.3 Main Results

We conduct extensive experiments to assess the performance of SAR, MAV, MAD, and EAG (ours) on BBH benchmark. Overall, our method obtains the best accuracy on 19/23 tasks and shows significant improvement on the unweighted average of 23 tasks (8% ↑). Next, we discuss the performance presented in Table 3 divided into three categories.

Mathematical Reasoning Some tasks in BBH require agents to perform mathematical computations (e.g. *Multi-Step Arithmetic*) or reasoning (e.g. *Word Sorting*). Among the tasks, *Multi-Step Arithmetic* is extremely challenging for LLMs (Suzgun et al., 2023), which leads to the poor performance of single-agent reasoning and multi-agent voting



Q: ((0+ 3 * 4 * 9) - (4 - -5 + -4 - 4)) = ?

Round 1, Agent 1:
Answer:113 ✗
Explanation: First, solve the multiplication within the parentheses:
3 * 4 * 9 = 108.
Then simplify the subtraction:
-4 - (-5) + (-4) - 4 = -4 + 5 - 4 - 4=-7.
Finally, subtract the second result from the first:
108 - (-7) = 108 + 7= 113.

Round 1, Agent 2:
Answer:109 ✗
Explanation: First, calculate the expression inside the parentheses:
(0 + 3* 4 * 9) - (4 - -5 + -4 - 4)
= (0 + 108) - (0 - 9) = 108 - -9 = 117.

Round 1, Agent 3:
Inconsistent Answering with Reasoning!
Answer:**105** ✗
Explanation: First simplifying the expression inside the parentheses,
-4 +-5 = -9,   -9 - 4 = -13.
Therefore, the expression becomes
(0 + 3 * 4 * 9) -(-13),
which simplifies to
(0 + 108) - (-13) = 108 + 13 = **121**.

Round 2, Agent 1:
Answer:113 ✗
Explanation: First, calculate the expression within the parentheses as done by the first agent:
(0 + 3 * 4 * 9) - (4 - -5 + -4 - 4)
= (0 + 108) - (0 - 9) =108 - (-9)
= 117. However, the second agent's reasoning also provides a valid approach. Simplify the subtraction within the parentheses:
-4 - (-5) + (-4) - 4= -4 + 5 - 4 - 4 =-7.
Thus, the final calculation is
108 - (-7) = 108 + 7 =113,
which would be the correct answer.

Round 2, Agent 2:
misleading phenomenon !
Answer:**113** ✗
Explanation: First, calculate the expression inside the parentheses:
(0 + 3 * 4 * 9) - (4 - -5 + -4 - 4)
= (0 +108) - (0 - 9)
= 108 - -9
= 117.
Additionally, following the reasoning from **the first agent's** solution,
the final calculation should be:
117 - 4 = 113.

Round 2, Agent 3:
Answer:**108** ✗
Explanation: First, let's simplify the expression inside the parentheses. 3* 4 * 9 = 108.
Now, let's simplify the subtraction inside the parentheses.
-4 -(-5) +(-4)- 4 = -4 + 5 - 4 - 4 = -3.
Substitute these values back into the original expression:
(0 + 108) - (-3) = 108 + 3 = 108.
Therefore, the updated answer is 108.

Explainer:
The question involves a mathematical expression that needs to be evaluated. The expression is as follows:
((0 + 3 * 4 * 9) -(4 - -5 + -4 - 4))
We need to follow the order of operations **(PEMDAS/BODMAS)** to solve this expression.

Analyzer:
To solve the given expression, we need to follow the order of operations **(PEMDAS/BODMAS)**, which stands for Parentheses, Exponents, Multiplication and Division (from left to right), and Addition and Subtraction (from left to right).
Given expression: (0 + 3 * 4 * 9) - (4 --5 + -4 - 4)
1. First, we need to perform the multiplication within the parentheses:
3 * 4 * 9 = 108
2.Next, we substitute the result back into the expression:
(0 + 108) - (4 - -5 + -4 - 4)
3. Now, we simplify the addition and subtraction within the parentheses:
(0 + 108) - (4 + 5 - 4 - 4) = 108 - 1 = 107
Therefore, the final answer to the expression
(0 + 3 * 4 * 9) -(4 - -5 + -4 - 4) is **107**.

Generator:
Answer: **107** ✓

Figure 6: Case study to illustrate the differences between MAD and our EAG method.

approaches. After the four rounds of debate, the performance of MAD exceeds multi-agent voting by 33% and may increase with additional rounds according to the Table 3, but the costs would also escalate. Our EAG approach lead to a further 36% enhancement compared to MAD. We attribute this to our method decomposing complex problems, enabling agents to collaborate on tasks that individual agents cannot accomplish independently.

Natural Language Understanding In some tasks within BBH, agents are required to possess ample world knowledge (e.g. *Movie Recommendation*) and accurately grasp the semantics of questions(e.g. *Date Understanding*). We observe that single-agent reasoning and multi-agent voting generally perform well in some tasks thanks to the vast reservoir of world knowledge in LLMs, while perform poorly in tasks involving misleading information in the questions, such as *Date Understanding*. The improvement of MAD on *Date Understanding* is also not substantial. Our EAG method achieves the 22% enhancement compared to the multi-agent voting and 13% enhancement compared to the MAD on *Date Understanding* task.

Scenario-based Question Answering Several tasks in BBH simulate a scene based on the real world and propose some related questions.(e.g. *Penguins in a Table* Compared to others, these tasks are closer to the real world, presenting challenge for LLMs. Consequently, single-agent reasoning and multi-agent voting perform moderately on such tasks. MAD and our EAG method both perform well on such tasks due to they are human-like collaborative approach. Moreover, our EAG method outperforms MAD on *Penguins in a Table* task by 10% and on *Reasoning about Colored Objects* by 20%.

4.4 Case Study

The performance comparison of MAD and EAG methods on *Multi-Step Arithmetic* task is shown in the Figure 6. In MAD approach, all agents generate wrong answers. *Agent 2* encounter *inconsistent answers with reasoning* error in Round 1 and is misled by *Agent 3* to generate the same erroneous answer. In contrast, explainer agent in our EAG method proposes the (PEMDAS/BODMAS) rule, which helps subsequent agents in generating the correct solving process and answer. See Figure 8 in the Appendix for more cases on *Movie Recommendation* task.

4.5 Ablation Study

We conducted ablation experiments to test the performance of *explainer*, *analyzer*, *generator* agent separately in the EAG framework on *Tracking Shuffled Objects (TSO)*, *Date Understanding (DU)*, *Penguins in a Table (PIA)* and *Geometric Shapes (GS)* tasks presented in Table 4.

Analyzer-Generator *w/o* Explainer The performance of EAG decrease after removing the explainer agent. We observed an 8% decrease in performance on the TSO task, which indicate that

Table 4: Ablation study to illustrate the effect of *Explainer*(E), *Analyzer*(A) and *Generator*(G) agent on some typical tasks in BBH.

E	A	G	Tasks			
			TSO	DU	PIA	GS
✗	✓	✓	25.8±1.8	71.3±1.0	70.3±1.0	22.3±0.3
✓	✗	✓	21.6±0.7	53.0±0.5	54.7±1.2	22.7±1.4
✗	✗	✓	23.7±1.1	47.3±0.5	49.3±0.7	17.3±0.3
✓	✓	✓	**34.4**±2.9	**72.0**±2.9	**71.3**±4.4	**31.7**±1.4

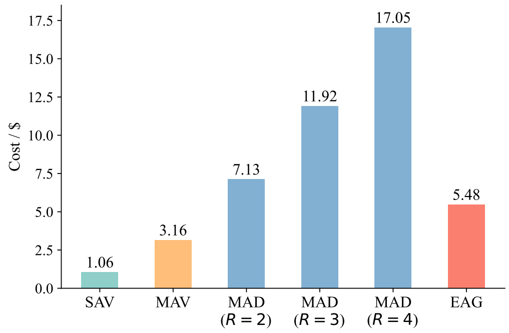


Figure 7: Cost comparison between different collaboration methods.

explainer agent plays a role in the EAG method.

Explainer-Generator *w/o* Analyzer Removing analyzer agent leads to a significant decrease in the performance of EAG. We observe an 10%-20% decrease in performance on the four tasks. This indicates that analyzer is the core agent of the architecture.

Only Generator Removing explainer and analyzer, the performance of EAG reaches the worst. In the TSO task, the performance of Only Generator is better than Explainer-Generator. We believe this is due to the information from explainer includes some redundant information that generator cannot comprehend after removing analyzer agent.

4.6 Cost Analysis

Through the cost comparison shown in Figure 7, we observe that the costs of MAD is the most expensive and increases with the debate rounds. In contrast, the costs of EAG is lower and more stable. Combining the performance presented in the Table 3, we find that the performance of multi agent voting almost on par with MAD with lower costs and the significant costs do not bring corresponding performance improvements to MAD. Our EAG method achieve the best average results with moderate costs.

5 Related Work

5.1 LLMs and Prompt Engineering

The emergence of Large Language Models (LLMs) has brought hope for achieving General Artificial Intelligence(GAI). Typically comprising hundreds of billions (or more) of parameters, exemplified by models such as GPT-3(Brown et al., 2020b), GLM(Zeng et al., 2023), Galactica(Taylor et al., 2022), and LLaMA(Touvron et al., 2023a), they often adhere to Scaling Law principles(Kaplan et al., 2020). To further unleash the LLMs' potential, prompt engineering(Liu et al., 2023) is extensively employed. However, these methods are designed specifically for individual large language models.

5.2 Multi-Agent Collaboration

Inspired by multi-agent settings, Du et al. (2023) has proposed a novel debate approach for large models , Liang et al. (2023) has introduced a Multi-Agent Debate (MAD) framework , wherein multiple agents articulate their arguments while a judge orchestrates the debate to reach a conclusive resolution. 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.Recently, Chen et al. (2023) has proposed ReConcile, which emulates a round table conference, enhancing collaborative reasoning among diverse LLM agents to achieve a more robust consensus.

6 Conclusion and limitations

This paper comprehensively analyzes the limitations of the MAD method on the BBH benchmark. By analyzing the erroneous samples of MAD on BBH, we summarize the types of errors and propose a novel multi-agent collaboration approach named EAG. By decomposing complex tasks, EAG fully leverages LLMs with moderate costs. Qualitative and quantitative experiments on BBH demonstrate the effectiveness of EAG.

The current design of EAG is fixed with three agents, which is not flexible enough. How to enable efficient collaboration among different types of multi-agents is a research topic deserving separate study.We will continue to refine EAG in the upcoming work according to these directions.

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.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of AAAI*.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Proceedings of NeurIPS*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of NeurIPS*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Proceedings of NeurIPS*.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through LLM negotiations. *arXiv preprint arXiv:2311.01876*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Proceedings of ACL*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2309.16609*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of NeurIPS*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Proceedings of EMNLP*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of NeurIPS*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *Proceedings of ICLR*.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. 2023. Solving math word problems via cooperative reasoning induced language models. In *Proceedings of ACL*.

A Appendix

A.1 The Comparison between MAD and MAV

We conduct accuracy and consistency changes of MAV and MAD experiments across different debate rounds and agent counts on BBH. The trend of data changes is shown in the Figure 2, with specific data presented in Table 5 and Table 6.

Table 5: Accuracy and Consistency under different rounds.

	Task	R=1	R=2	R=3
Accuracy	Navigate_Voting	63.00	63.00	63.00
	Navigate_MAD	63.00	58.00	55.70
	Snarks_Voting	75.00	75.00	75.00
	Snarks_MAD	75.00	72.70	73.30
	Average_Voting	48.44	48.44	48.44
	Average_MAD	48.44	50.13	50.22
Consistency	Navigate_Voting	79.56	79.56	79.56
	Navigate_MAD	79.56	84.22	88.56
	Snarks_Voting	90.00	90.00	90.00
	Snarks_MAD	90.00	97.44	96.22
	Average_Voting	76.70	76.70	76.70
	Average_MAD	76.70	85.57	89.60

Table 6: Accuracy and Consistency under different agents.

	Task	A=3	A=4	A=5
Accuracy	Navigate_Voting	63.00	60.00	61.33
	Navigate_MAD	58.00	57.33	56.00
	Snarks_Voting	75.00	78.33	77.33
	Snarks_MAD	72.67	79.00	75.00
	Average_Voting	48.44	48.60	48.51
	Average_MAD	50.15	50.27	50.00
Consistency	Navigate_Voting	79.56	75.92	76.27
	Navigate_MAD	84.22	82.83	84.53
	Snarks_Voting	90.00	89.00	87.60
	Snarks_MAD	94.44	94.08	94.26
	Average_Voting	76.70	74.30	73.56
	Average_MAD	85.57	84.86	85.52

A.2 The Error Examples of Single-Agent on BBH

We conduct a statistical analysis of the errors in MAD on BBH and divided the erroneous samples into three categories, including inaccurate question understanding, faulty reasoning steps and inconsistent answers with reasoning. The typical example in each category is presented in Table 9.

A.3 The Prompt of EAG framework

Our EAG framework contains Explainer, Analyzer and Generater agents. The prompts of the different agents are detailed as follows:

Explainer Prompt:
System: You are collaborating with other agents to answer the question. The process includes explaining, analyzing, and answering the question. Your task is to explain the question.
Question:$\{q\}$
User: Please review and share your understanding of the question to assist the following agents in improving their problem-solving abilities. Ensure you focus on essential information while filtering out irrelevant details.

Analyzer Prompt:
System: You are collaborating with other agents to answer the question. The process includes explaining, analyzing, and answering the question. Your task is to analyze the question.
Question:$\{q\}$
Explained information:$\{y_{A_e}\}$
User: Please review the question and the understanding from the previous agent, and provide a comprehensive reasoning process to solve the problem. Ensure clarity and detail in your thought process to effectively guide the following agent in generating the final answer.

Generator Prompt:
System: You are collaborating with other agents to answer the question. The process includes explaining, analyzing, and answering the question. Your task is to generate the question.
Question:$\{q\}$
Explained information:$\{y_{A_e}\}$
Question analysis: $\{y_{A_a}\}$
User: Please review the question and the thoughts of previous agents, and provide the final answer to the question. Ensure to follow this format: “Answer: (A|B|C|D|E|F|G|H|I|J)”.

A.4 Dataset Details

We conduct experiments on BBH benchmark, which contains 23 complex reasoning tasks including mathematical calculations, logical reasoning, common sense questions, multilingual problems, geometry, etc. The statistics of BBH are shown in Table 7. The description of each task in BBH is presented in Table 10.

Table 7: Statistics of datasets used in the BIG-Bench Hard benchmark, categorized by task type.

Category	Number of Tasks	Queries
Math	6	1,500
NLU	10	2,365
ScenarioQA	7	2,646
All	23	6511

A.5 Result Format Prompt

The generation of LLMs is stochastic. To better capture the final answers of LLMs in a zero-prompt setting, we constrain the outputs of LLMs. We design different prompts presented in Table 8 based on three types of questions, including multiple-choice question, true/false question and text generation question.

Table 8: The prompt of the result format.

Types of Question	Prompt
Multiple Choice	Please use the format like “Answer: (A|B|C|D|E|F|G|H|I|J)”
True / False	Please answer true or false with the format like “Answer: .” || Please answer valid or invalid with the format like “Answer:.”|| Please answer yes or no with the format like “Answer: .”
Text Generation	Please use the format like “Answer:”.

Table 9: The error examples of single-agent on BBH.

Inaccurate question understanding

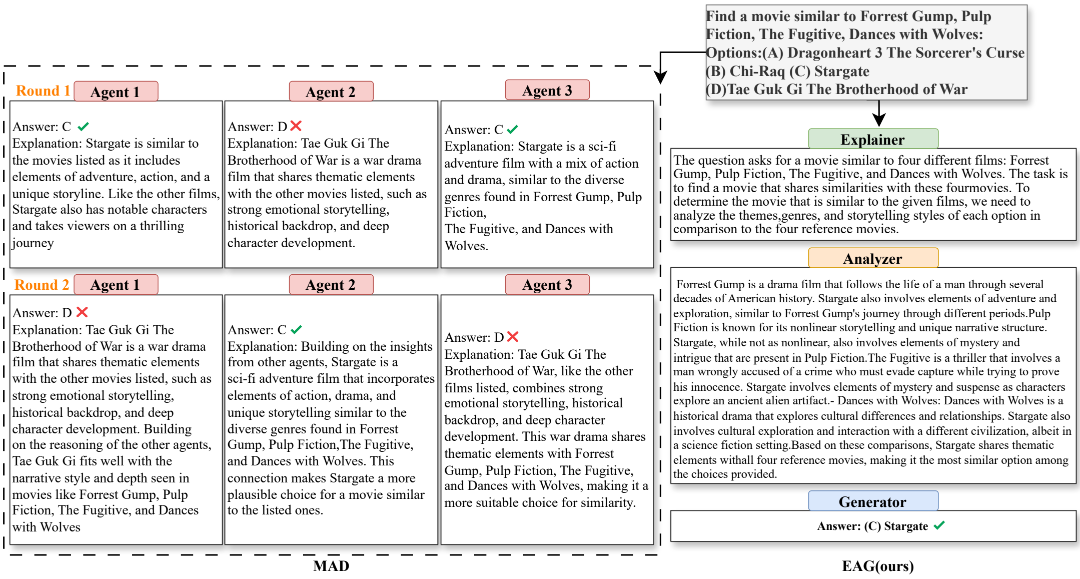
Question: Today is the second day of the third month of 1966. What is the date tomorrow in MM/DD/YYYY? Options:
(A) 03/03/1966 (B) 03/24/1966 (C) 03/04/1966 (D) 02/28/1966 (E) 03/02/1966
Answer: (C)
Explanation: Today is March 2, 1966. Tomorrow will be March 3, 1966, so the correct option is 03/04/1966
GrounfTruth: (A).

Faulty reasoning steps

Question: Alice, Bob, and Claire are dancers at a square dance. At the start of a song, they each have a partner: Alice is dancing with Rodrigo, Bob is dancing with Jamie, and Claire is dancing with Sam. Throughout the song, the dancers often trade partners. First, Bob and Alice switch partners. Then, Bob and Claire switch partners. Finally, Alice and Claire switch partners. At the end of the dance, Alice is dancing with Options:
(A) Rodrigo (B) Jamie (C) Sam
Answer: (C)
Explanation: Since Alice starts with Rodrigo and then switches partners with Bob and Claire throughout the song, at the end of the dance, Alice would be dancing with Sam.
GrounfTruth: (A).

Inconsistent answers with reasoning

Question: Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. What is the average height of the penguins? Options:
(A) 60 (B) 65 (C) 70 (D) 75 (E) 80
Answer: (C)
Explanation: To find the average height of the penguins, we add up all the heights (50 + 80 + 60 + 70 = 260) and divide by the total number of penguins (4), which equals 65.
GrounfTruth: (B).



Find a movie similar to Forrest Gump, Pulp Fiction, The Fugitive, Dances with Wolves: Options:(A) Dragonheart 3 The Sorcerer's Curse (B) Chi-Raq (C) Stargate (D)Tae Guk Gi The Brotherhood of War

Round 1, Agent 1: Answer: C ✓
Explanation: Stargate is similar to the movies listed as it includes elements of adventure, action, and a unique storyline. Like the other films, Stargate also has notable characters and takes viewers on a thrilling journey

Round 1, Agent 2: Answer: D ✗
Explanation: Tae Guk Gi The Brotherhood of War is a war drama film that shares thematic elements with the other movies listed, such as strong emotional storytelling, historical backdrop, and deep character development.

Round 1, Agent 3: Answer: C ✓
Explanation: Stargate is a sci-fi adventure film with a mix of action and drama, similar to the diverse genres found in Forrest Gump, Pulp Fiction,
The Fugitive, and Dances with Wolves.

Round 2, Agent 1: Answer: D ✗
Explanation: Tae Guk Gi The Brotherhood of War is a war drama film that shares thematic elements with the other movies listed, such as strong emotional storytelling, historical backdrop, and deep character development. Building on the reasoning of the other agents, Tae Guk Gi fits well with the narrative style and depth seen in movies like Forrest Gump, Pulp Fiction, The Fugitive, and Dances with Wolves

Round 2, Agent 2: Answer: C ✓
Explanation: Building on the insights from other agents, Stargate is a sci-fi adventure film that incorporates elements of action, drama, and unique storytelling similar to the diverse genres found in Forrest Gump, Pulp Fiction,The Fugitive, and Dances with Wolves. This connection makes Stargate a more plausible choice for a movie similar to the listed ones.

Round 2, Agent 3: Answer: D ✗
Explanation: Tae Guk Gi The Brotherhood of War, like the other films listed, combines strong emotional storytelling, historical backdrop, and deep character development. This war drama shares thematic elements with Forrest Gump, Pulp Fiction, The Fugitive, and Dances with Wolves, making it a more suitable choice for similarity.

Explainer: The question asks for a movie similar to four different films: Forrest Gump, Pulp Fiction, The Fugitive, and Dances with Wolves. The task is to find a movie that shares similarities with these fourmovies. To determine the movie that is similar to the given films, we need to analyze the themes,genres, and storytelling styles of each option in comparison to the four reference movies.

Analyzer: Forrest Gump is a drama film that follows the life of a man through several decades of American history. Stargate also involves elements of adventure and exploration, similar to Forrest Gump's journey through different periods.Pulp Fiction is known for its nonlinear storytelling and unique narrative structure. Stargate, while not as nonlinear, also involves elements of mystery and intrigue that are present in Pulp Fiction.The Fugitive is a thriller that involves a man wrongly accused of a crime who must evade capture while trying to prove his innocence. Stargate involves elements of mystery and suspense as characters explore an ancient alien artifact.- Dances with Wolves: Dances with Wolves is a historical drama that explores cultural differences and relationships. Stargate also involves cultural exploration and interaction with a different civilization, albeit in a science fiction setting.Based on these comparisons, Stargate shares thematic elements withall four reference movies, making it the most similar option among the choices provided.

Generator: **Answer: (C) Stargate** ✓

Figure 8: Case study to illustrate the differences between MAD and our EAG method.

Table 10: Some BBH task description.

Task Name	Task Description
Boolean Expressions	Evaluate the truth value of a random Boolean expression consisting of Boolean constants and basic Boolean operators.
Causal Judgment	Given a short story, determine how a typical person would answer a causal question about the story.
Date Understanding	Given a small set of sentences about a particular date, answer the provided question.
Geometric Shapes	Given a full SVG path element containing multiple commands, determine the geometric shape that would be generated if one were to execute the full path element.
Hyperbaton (Adjective Ordering)	Given two English-language sentences, determine the one with the correct adjective order.
Logical Deduction	Deduce the order of a sequence of objects based on the clues and information about their spacial relationships and placements.
Movie Recommendation	Given a list of movies a user might have watched and liked, recommend a new, relevant movie to the user out of the four potential choices user might have.
Multi-Step Arithmetic	Solve multi-step equations involving basic arithmetic operations (addition, subtraction, multiplication, and division).
Navigate	Given a series of navigation steps to an agent, determine whether the agent would end up back at its initial starting point.
Object Counting	Given a collection of possessions that a person has along with their quantities, determine the number of a certain object/item class.
Penguins in a Table	Given a unique table of penguins, answer a question about the attributes of the penguins.
Reasoning about Colored Objects	Given a context, answer a simple question about the color of an object on a surface.
Web of Lies	Evaluate the truth value of a random Boolean function expressed as a natural-language word problem.
Word Sorting	Given a list of words, sort them lexicographically.