MAPS: Multi-Agent Personality Shaping for Collaborative Reasoning

Jian Zhang[*1,2], **Zhiyuan Wang**[*1,3], **Zhangqi Wang**[*1,3], **Fangzhi Xu**[1,3], **Qika Lin**[4], **Lingling Zhang**[1,3], **Rui Mao**[5], **Erik Cambria**[5], **Jun Liu**[1,2†]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]MOE KLINNS Lab, Xi'an Jiaotong University, China
[3]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, China
[4]Saw Swee Hock School of Public Health, National University of Singapore, Singapore
[5]College of Computing and Data Science, Nanyang Technological University, Singapore
zhangjian062422@stu.xjtu.edu.cn, liukeen@xjtu.edu.cn

Abstract

Collaborative reasoning with multiple agents offers the potential for more robust and diverse problem-solving. However, existing approaches often suffer from homogeneous agent behaviors and lack of reflective and rethinking capabilities. We propose **M**ulti-**A**gent **P**ersonality **S**haping ((**MAPS**), a novel framework that enhances reasoning through agent diversity and internal critique. Inspired by the Big Five personality theory, MAPS assigns distinct personality traits to individual agents, shaping their reasoning styles and promoting heterogeneous collaboration. To enable deeper and more adaptive reasoning, MAPS introduces a *Critic* agent that reflects on intermediate outputs, revisits flawed steps, and guides iterative refinement. This integration of personality-driven agent design and structured collaboration improves both reasoning depth and flexibility. Empirical evaluations across three benchmarks demonstrate the strong performance of MAPS, with further analysis confirming its generalizability across different large language models and validating the benefits of multi-agent collaboration.

Code — https://github.com/exoskeletonzj/MAPS

Introduction

Solving complex reasoning problems (Bhattacharya et al. 2024; Li et al. 2024c; He et al. 2024) often requires more than just accurate perception and factual recall—it demands nuanced interpretation, multi-step inference, and the ability to adapt when initial attempts fail (Fu et al. 2024; Li et al. 2024a). While large language models (LLMs) demonstrate promising capabilities in solving such problems, they frequently fall short in scenarios requiring sustained reasoning, internal verification, and flexible strategy revision (Anand et al. 2024; Alasadi and Baiz 2024; Gao et al. 2024; Wang et al. 2024b; Xu et al. 2025). A key challenge lies in how to effectively combine diverse reasoning strategies and incorporate mechanisms for intermediate reflection. Prior work has explored both single-agent solutions and simple collaborative setups (e.g., paired discussion or voting) (Kaesberg et al. 2025), yet these approaches often suffer from rigid behaviors and limited capacity for self-correction. This raises an important research question: *how can we enhance the depth and adaptability of reasoning by enabling more flexible and reflective solution processes?* Figure 1 illustrates a representative scenario that embodies these challenges, where successful problem solving requires interpreting multimodal input and applying domain-specific reasoning.

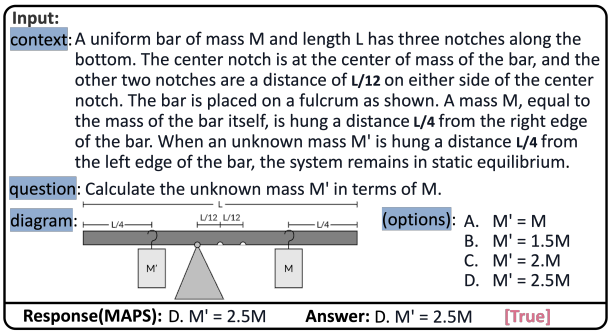

Figure 1: An example of a multimodal scientific multiple-choice problem. The correct answer is derived based on the reasoning over inputs that include context, question, and diagram.

As illustrated in Figure 2, existing methods (Wang et al. 2024a; Landau, Páez, and Bordeianu 2024; Hardiansyah et al. 2024; Zhang et al. 2024; Caffagni et al. 2024; Qiu, Yuan, and Lam 2024) for complex reasoning problems often adopt single-agent solutions or simple two-agent collaborations. Although effective to some extent, these setups suffer from **homogeneous agent behaviors**—reasoning steps tend to repeat similar patterns, leading to redundancy and premature convergence. The lack of diversity limits exploration and reduces the chance of identifying alternative perspectives or correcting errors, especially in multi-turn settings where roles and strategies remain undifferentiated.

Another issue is the **lack of reflective and rethinking capabilities** in existing approaches. As shown in Figure 2, the interaction between two agents is often linear and lacks mechanisms for feedback or revision. Even with multiple

[*]These authors contributed equally.
[†]Corresponding Author

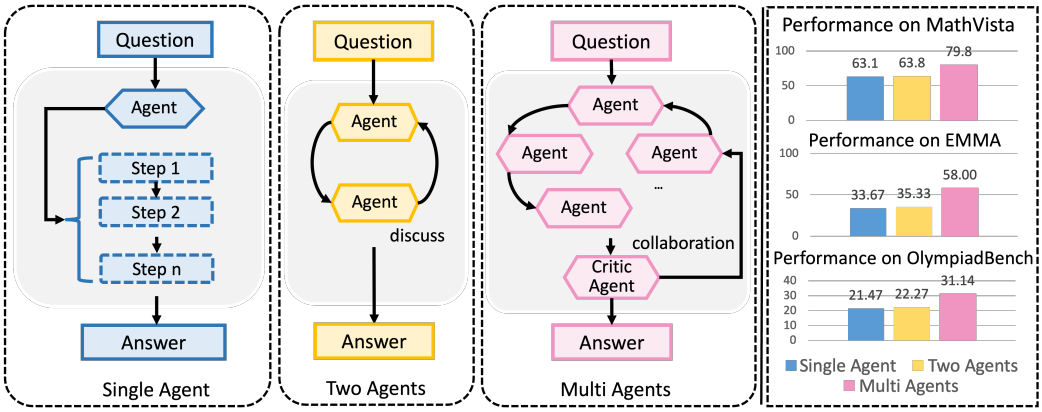

Figure 2: Comparison of reasoning paradigms. Single-agent and two-agent approaches offer limited adaptability. MAPS enables dynamic collaborative reasoning. Right: Built on GPT-4o, MAPS achieves the best performance across three benchmarks.

turns, agents rarely revisit earlier reasoning or correct initial misconceptions. In contrast, human reasoning is inherently iterative: People reflect, reassess, and adjust their thinking over time. Without structured reflection, current methods risk premature convergence and fail to recover from early-stage errors.

To address these two issues, we propose **MAPS** (**M**ulti-**A**gent **P**ersonality **S**haping), a collaborative reasoning framework that enhances both diversity and adaptability in complex problem solving. Inspired by the Big Five personality theory (Almagor, Tellegen, and Waller 1995; Benet and Waller 1995; Simms 2007), MAPS shapes the reasoning behaviors of a set of role-specialized agents through distinct personality traits, promoting heterogeneous collaboration and mitigating behavioral homogeneity. To enable reflective thinking and iterative refinement, MAPS further introduces a *Critic* agent that revisits intermediate outputs, identifies flawed steps, and provides structured feedback. This integration of personality-guided agent design and internal critique supports deeper, more flexible reasoning aligned with human cognitive processes.

As shown in Figure 3, the *interpreter* (Openness) explores information from multiple angles, the *aligner* (Agreeableness) reconciles visual and textual cues, the *scholar* (Conscientiousness) ensures precision and factual grounding, the *solver* (Extraversion) drives goal-oriented conclusions, and the *critic* (Neuroticism) questions assumptions and detects flaws. The *critic* is further inspired by Socratic questioning (Elder and Paul 1998; Paul and Elder 2019), providing reflective feedback throughout the multi-stage reasoning process. Together, these roles enable a structured yet flexible collaboration that supports deeper and more reliable problem solving.

We conduct extensive experiments on three challenging benchmarks: MathVista (Lu et al. 2023), OlympiadBench (He et al. 2024), and EMMA (Hao et al. 2025). These datasets cover a wide range of complex reasoning tasks. MAPS consistently outperforms baseline methods across all datasets, confirming its effectiveness in enhancing both reasoning depth and adaptability.

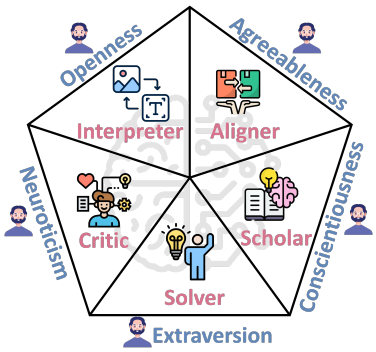

Figure 3: The corresponding relation between the Big Five Personality theory and the five function-specific agents.

We further evaluate MAPS under different base LLMs, and results consistently show its superiority across model backbones. Additional analyses examine the impact of the feedback mechanism and the overall efficiency of the framework.

Our main contributions are as follows:

- We propose **MAPS**, a multi-agent reasoning framework. To the best of our knowledge, this is the first work that incorporates personality shaping based on the Big Five theory into collaborative reasoning.
- MAPS addresses two key challenges in existing methods: behavioral homogeneity and lack of reflection. It assigns distinct personality traits to agents and introduces a *Critic* inspired by Socratic questioning.
- We conduct extensive experiments on three scientific reasoning benchmarks. MAPS achieves consistent performance gains (up to 15.84%) and generalizes well across different tasks and base models.

Methodology

This section introduces MAPS in four key components: *Preliminaries*, *Agentic Interaction Logic*, the *Four-Step Reasoning* process, and the *Critic and Feedback* mechanism.

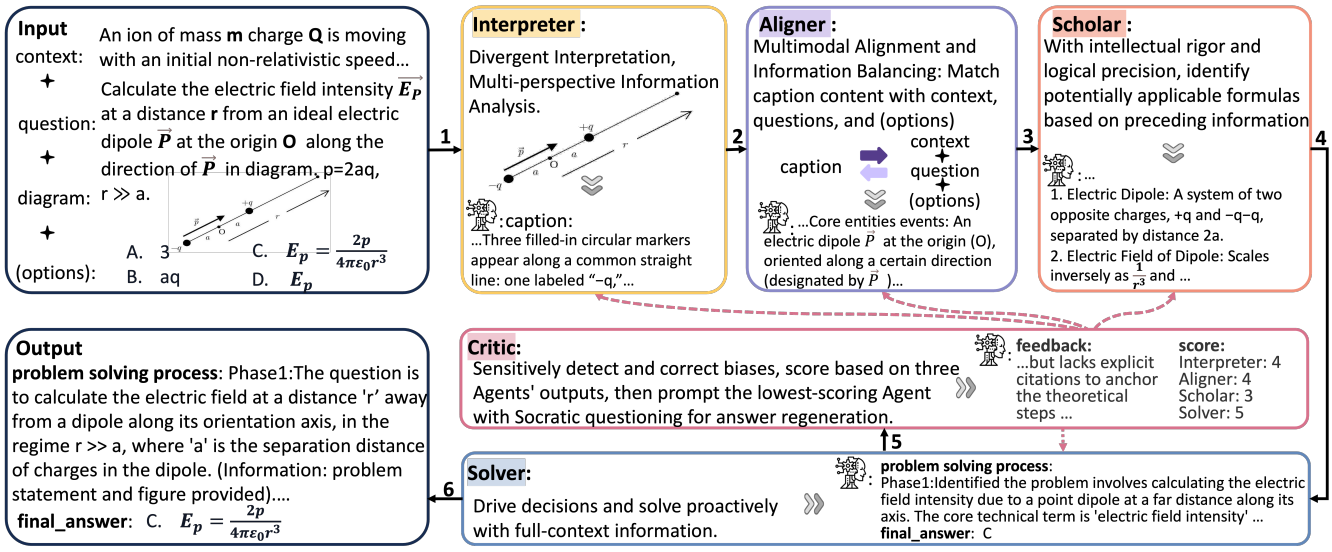

Figure 4: The overall architecture of MAPS. The framework consists of five functional agents inspired by the Big Five personality theory. The core reasoning process is carried out by four specialized agents—*Interpreter*, *Aligner*, *Scholar*, and *Solver*—each responsible for a distinct stage in solving complex reasoning problems. Finally, the *Critic* agent provides reflective feedback and correction to enhance accuracy and interpretability.

Preliminaries

Task Definition. We define the complex reasoning task as a mapping from structured multimodal inputs to a target answer space. Let $\mathcal{D}$, $\mathcal{C}$, and $\mathcal{Q}$ denote the input spaces of diagrams, contexts, and questions, respectively, and let $\mathcal{A}$ denote the answer space. Each instance is a triplet $(d_i, c_i, q_i) \in \mathcal{D} \times \mathcal{C} \times \mathcal{Q}$, and the goal is to predict the corresponding answer $a_i \in \mathcal{A}$.

The task is thus modeled as a function:

$$\mathcal{M} : \mathcal{D} \times \mathcal{C} \times \mathcal{Q} \to \mathcal{A}, \quad a_i = \mathcal{M}(d_i, c_i, q_i), \tag{1}$$

where $\mathcal{M}$ denotes the reasoning system responsible for processing visual-textual input and generating the output.

Agent-Based Modeling. In MAPS, the reasoning process $\mathcal{M}$ is decomposed into a sequence of collaborative stages executed by a set of role-specialized agents $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_K\}$, where each agent $\mathcal{A}_k$ performs a specific function conditioned on both the original input and intermediate reasoning states.

Let $\mathcal{S}_0 = (d_i, c_i, q_i)$ be the initial input state. The overall reasoning unfolds as a staged transformation:

$$\mathcal{S}_k = \mathcal{A}_k(\mathcal{S}_{k-1}), \quad \text{for } k = 1, 2, \ldots, K, \tag{2}$$

where $\mathcal{S}_k$ denotes the intermediate reasoning state after the k-th agent's operation. The final answer is extracted as $a_i = \text{Extract}(\mathcal{S}_K)$. This formulation reflects the staged reasoning in MAPS, where specialized agents collaboratively refine the solution.

Agents Interaction Logic

As illustrated in Algorithm 1, this section introduces the interaction logic among the five personality-driven agents in MAPS. Complex problem reasoning requires multimodal semantic integration and structured inference over multiple steps. MAPS models this process as a functional composition of role-specialized reasoning agents, each shaped by a distinct personality trait.

Let $\mathbf{x} = (d_i, c_i, q_i)$ be the input, and let $\mathbf{p}_k \in \mathbb{R}^m$ denote the personality embedding associated with agent $\mathcal{A}_k$. The entire collaborative reasoning process is represented as:

$$a_i = \mathcal{F}(\mathbf{x}; \mathbf{p}_1, \ldots, \mathbf{p}_4) = \mathcal{A}_4 \circ \mathcal{A}_3 \circ \mathcal{A}_2 \circ \mathcal{A}_1(\mathbf{x}), \tag{3}$$

where each agent $\mathcal{A}_k(\cdot; \mathbf{p}_k)$ executes its stage conditioned on its personality vector $\mathbf{p}_k$, contributing a unique reasoning perspective.

After the initial reasoning trajectory $\mathcal{T} = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$ is generated, the *Critic* agent performs a reflective evaluation by applying a feedback function:

$$\mathbf{f} = \texttt{Reflect}(\mathcal{T}), \quad \mathbf{f} = \{f_1, f_2, f_3, f_4\} \in \mathbb{R}^4, \tag{4}$$

where each score f_k reflects the confidence of the k-th stage. If any $f_k < \tau$, the corresponding stage is revised. This feedback loop complements the forward reasoning path, allowing MAPS to emulate deliberative human cognition through structured collaboration and critique.

Proposition 1 (Monotonic Free-Energy Descent). *For the free energy* $F^{(t)} = \mathbb{E}_{q^{(t)}}[-\log p(\mathbf{x}, a_i \mid \theta)] + \mathrm{KL}\big(q^{(t)} \| p\big)$, *each Critic-triggered update satisfies*

$$F^{(t+1)} \leq F^{(t)}. \tag{5}$$

Thus the MAPS iteration produces a non-increasing free-energy sequence that converges to a stationary point.

Algorithm 1: MAPS Collaborative Reasoning with Reflective Feedback

Input: Diagram d_i, Context c_i, Question q_i, Personality vectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_4\}$
Initialize: $\mathcal{S}_0 = (d_i, c_i, q_i), \quad t \leftarrow 0$
repeat
  Interpreter: $p_i = \mathcal{A}_1(\mathcal{S}_0; \mathbf{p}_1)$
  Aligner: $l_i = \mathcal{A}_2(p_i, c_i, q_i; \mathbf{p}_2)$
  Scholar: $s_i = \mathcal{A}_3(l_i, p_i, c_i, q_i; \mathbf{p}_3)$
  Solver: $a_i = \mathcal{A}_4(s_i, l_i, p_i; \mathbf{p}_4)$
  Reasoning Trajectory: $\mathcal{T}^{(t)} = \{p_i, l_i, s_i, a_i\}$
  Critic: $\mathbf{f}^{(t)} = \texttt{Reflect}(\mathcal{T}^{(t)}), \quad f_k \in [0, 1]$
  if all $f_k \geq \tau$ **then**
    break
  else
    Identify stage $k^* = \arg\min_k f_k$
    Rerun agent $\mathcal{A}_{k^*}$ with updated input
    $t \leftarrow t + 1$
  end if
until convergence
return Final answer a_i

Four-Step Reasoning

Given input $\mathbf{x} = (d_i, c_i, q_i)$, MAPS conducts a structured reasoning process across four personality-driven agents: *Interpreter*, *Aligner*, *Scholar*, and *Solver*. Each agent $\mathcal{A}_k$ is parameterized by a personality embedding $\mathbf{p}_k \in \mathbb{R}^m$, shaping its reasoning style and focus. The entire inference pipeline can be expressed as a function composition:

$$a_i = \mathcal{A}_4 \circ \mathcal{A}_3 \circ \mathcal{A}_2 \circ \mathcal{A}_1(\mathbf{x}; \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4), \tag{6}$$

where each intermediate state $\mathcal{S}_k$ is defined recursively as $\mathcal{S}_k = \mathcal{A}_k(\mathcal{S}_{k-1}; \mathbf{p}_k)$, with $\mathcal{S}_0 = \mathbf{x}$.

Interpreter. The *Interpreter* agent aims to extract structured visual semantics from diagram d_i, translating them into a caption representation p_i that can be consumed by downstream language agents. Let $\phi_{\text{vis}}(d_i)$ be a visual encoder, and $\psi_{\text{lang}}(\cdot)$ a caption generator. Then the agent performs:

$$p_i = \mathcal{A}_1(d_i; \mathbf{p}_1) = \psi_{\text{lang}}\big(\phi_{\text{vis}}(d_i) + W_1 \mathbf{p}_1\big), \tag{7}$$

where $W_1 \in \mathbb{R}^{d \times m}$ projects the personality embedding into the encoder space, modulating its attention to attributes like spatiality, color, or topology.

Aligner. To resolve semantic mismatches across modalities, the *Aligner* fuses the interpreted diagram caption p_i with the textual context c_i and question q_i. The process outputs an alignment representation l_i, optimized to preserve shared semantics and suppress modality conflict. Formally,

$$l_i = \mathcal{A}_2(p_i, c_i, q_i; \mathbf{p}_2) = \mathrm{CrossFuse}\big(p_i, c_i, q_i; \mathbf{p}_2\big), \tag{8}$$

where $\mathrm{CrossFuse}(\cdot)$ denotes a multi-head attention-based fusion operator, adaptively weighted by $\mathbf{p}_2$ to emphasize visual or linguistic cues based on agent bias.

Scholar. While l_i captures semantic consistency, complex reasoning often requires external knowledge supplementation. The *Scholar* agent retrieves and integrates domain-specific knowledge $\mathcal{K}(l_i)$, such as physics principles or mathematical theorems. We define:

$$s_i = \mathcal{A}_3(l_i, p_i, c_i, q_i; \mathbf{p}_3) = \mathrm{KnowAug}\big(l_i, \mathcal{K}(l_i); \mathbf{p}_3\big), \tag{9}$$

where $\mathrm{KnowAug}(\cdot)$ augments contextual embeddings with retrieved tuples $\mathcal{K}(l_i)$ from a structured knowledge memory, and $\mathbf{p}_3$ biases the agent toward formal rigor or heuristic reasoning.

Solver. The *Solver* agent aggregates all upstream outputs and executes logical composition to generate the final answer a_i. Let $\mathcal{H}_i = \{p_i, l_i, s_i\}$ be the hybrid reasoning state. The solver computes:

$$a_i = \mathcal{A}_4(\mathcal{H}_i; \mathbf{p}_4) = \mathrm{Deduct}\big(p_i, l_i, s_i; \mathbf{p}_4\big), \tag{10}$$

where $\mathrm{Deduct}(\cdot)$ is a constrained generation module that synthesizes the inputs under logical, numerical, or symbolic rules. The final prediction $a_i \in \mathcal{A}$ may be a selected option or a free-form answer.

Together, these four stages construct a trajectory $\mathcal{T} = \{p_i, l_i, s_i, a_i\}$, on which the *Critic* agent operates for reflection and feedback.

Proposition 2 (Collaborative Information Bottleneck). *Let $\mathcal{S}_k$ be the intermediate output of the k-th agent, given input $\mathbf{x} = (d_i, c_i, q_i)$ and target answer a_i. Then the MAPS reasoning process optimizes*

$$\min \sum_{k=1}^{4} I(\mathbf{x}; \mathcal{S}_k) \quad \text{s.t.} \quad I(\mathcal{S}_k; a_i) \geq \varepsilon, \tag{11}$$

where $I(\cdot;\cdot)$ denotes mutual information. The Critic monitors constraint violations and reactivates stages with insufficient task-relevant information.

Critic and Feedback

The *Critic* agent evaluates the internal reasoning trajectory $\mathcal{T} = \{p_i, l_i, s_i, a_i\}$ without relying on ground-truth answers. Inspired by Socratic questioning, it examines each stage's logic and justification to identify flawed assumptions and incomplete inferences.

We define the feedback vector as:

$$\mathbf{f}_i = \mathcal{M}_{\text{crit}}(p_i, l_i, s_i, a_i), \quad \mathbf{f}_i \in [0, 1]^4, \tag{12}$$

where each element $f_i^{(k)}$ represents the Critic's confidence in the correctness and completeness of stage k.

The weakest stage is selected by:

$$k^* = \arg\min_k f_i^{(k)}, \quad \text{if} \quad f_i^{(k^*)} < \tau, \tag{13}$$

which triggers a targeted revision. This reflection-driven loop promotes iterative self-correction and deepens reasoning reliability.

Models	CoT	MathVista			EMMA				OlympiadBench						Avg.
		Gen.	Math	Avg.	Math	Phy.	Chem.	Avg.	MECO	MZCE	MZCO	PECO	PZCE	Avg.	
Random Choice	-	26.09	22.78	24.30	13.00	23.00	27.00	21.00	0.67	0.33	0.00	1.75	0.33	0.87	16.06
Human Expert	-	56.09	55.74	55.90	75.00	64.50	86.00	75.17	48.00	34.67	30.36	54.17	12.33	37.80	52.73
Claude 3.5 Sonnet	-	68.04	63.15	65.40	23.00	34.00	44.00	33.67	20.67	13.00	10.71	10.75	14.00	13.23	37.43
Gemini 2.0 Flash	-	70.65	70.93	70.80	20.00	40.00	36.00	32.00	8.00	5.67	7.14	3.07	7.00	5.39	36.06
GPT-4o	-	65.22	61.30	63.10	30.00	38.00	33.00	33.67	23.33	20.33	19.64	22.15	21.00	21.47	39.41
Qwen2.5-VL-72B	-	70.65	67.41	68.90	42.00	42.00	38.00	40.67	18.00	12.33	5.36	7.24	3.67	8.80	39.45
InternVL2.5-8B-MPO	-	64.78	60.74	62.60	30.00	40.00	38.00	36.00	10.67	6.67	10.71	1.10	0.67	3.88	34.16
LLaVA-Onevision-72B	-	62.83	58.52	60.50	25.00	32.00	24.00	27.00	6.67	7.33	3.57	3.29	9.67	6.18	31.23
Claude 3.5 Sonnet	✓	71.74	64.26	67.70	30.00	38.00	41.00	36.33	24.00	11.00	16.07	12.72	10.33	13.23	39.09
Gemini 2.0 Flash	✓	70.22	75.56	73.10	24.00	41.00	36.00	33.67	12.67	6.33	3.57	4.61	2.33	5.39	37.38
GPT-4o	✓	65.22	62.59	63.80	27.00	44.00	35.00	35.33	25.33	21.67	12.50	24.12	20.33	22.27	40.47
Qwen2.5-VL-72B	✓	71.09	77.96	74.80	38.00	36.00	37.00	37.00	23.33	13.00	10.71	8.11	1.33	9.59	40.46
InternVL2.5-8B-MPO	✓	60.87	67.41	64.40	31.00	36.00	24.00	30.33	12.00	8.33	1.79	2.85	0.99	4.75	33.16
LLaVA-Onevision-72B	✓	71.09	64.44	67.50	23.00	26.00	23.00	24.00	11.33	8.67	5.36	4.82	3.33	6.18	32.56
MAPS (GPT-4o$_{base}$)	**-**	**75.87**	**83.15**	**79.80**	**52.00**	**71.00**	**51.00**	**58.00**	**46.00**	**30.33**	**32.14**	**28.51**	**28.33**	**31.14**	**56.31**

Table 1: Performance across 10 subtasks. Gen. = General (MathVista), Phy./Chem. = Physics/Chemistry (EMMA), MECO/ MZCO/ MZCE = English/Chinese COMP & CEE Math (OlympiadBench), PECO/ PZCE = English/Chinese Physics (OlympiadBench).

Experiments

Datasets and Baselines

We evaluate on three benchmarks for complex problem reasoning: MathVista, OlympiadBench, and EMMA. MathVista covers math and general science, OlympiadBench focuses on high-level math and physics, and EMMA includes math, physics, and chemistry.

We use GPT-4o (Achiam et al. 2023) as the base model. For comparison, we include leading multimodal large language models (MLLMs), both proprietary and open-source, tested under both direct and Chain-of-Thought (CoT) settings.

Main Results

MAPS achieves a new state-of-the-art (SOTA), surpassing human-level performance for the first time. As shown in Table 1, MAPS outperforms previous SOTA models by 15.84% and exceeds human expert performance by 3.58% across all tasks, highlighting its strength in solving complex multimodal problems. MAPS demonstrates robust performance across mathematical, physical, chemical, and general tasks, showcasing strong interdisciplinary reasoning. Its multi-agent design, based on the Big Five personality theory, enables effective collaboration and contributes to the SOTA results. The system excels in multimodal semantic integration and multi-step reasoning by jointly leveraging diagrams, contexts, and questions. Furthermore, the *Critic* agent applies Socratic feedback to refine responses, enhancing both accuracy and reliability on challenging tasks.

MAPS exhibits strong adaptability and robustness across diverse reasoning tasks. The evaluation datasets span a wide range of question types, modalities, and reasoning difficulties. MathVista features judgment, multiple-choice, and open-ended fill-in-the-blank questions with varied answer formats, requiring accurate intent understanding and response generation. OlympiadBench emphasizes challenging open-ended problems demanding multi-step symbolic reasoning, where small errors can lead to divergent outcomes. EMMA introduces multimodal complexity with diagrams embedded in both questions and answer choices. Through feedback-driven multi-agent collaboration and Socratic questioning, MAPS effectively handles these challenges, achieving SOTA and demonstrating strong generalization across heterogeneous reasoning scenarios.

Analysis of Critic Agent

Critic schema includes scoring and Socratic feedback. Each reasoning step is rated from 0–5, guiding whether the system should backtrack or proceed. Feedback is heuristic, encouraging rethinking rather than offering direct answers. The *Critic* uses these scores to decide whether to regenerate specific steps, ensuring robustness in the final answer.

The Critic enhances reasoning via Socratic feedback without using gold labels. As shown in Figure 5, the top illustrates the feedback schema, while the bottom shows feedback proportions across three datasets. The *Critic* prompts reflection and correction by encouraging agents to question assumptions rather than passively accept reasoning.

Solver receives the most feedback in EMMA and OlympiadBench. As shown in the lower half of Figure 5, feedback varies by dataset. For MathVista, most steps need no regeneration, aligning with our superior 5.0% SOTA im-

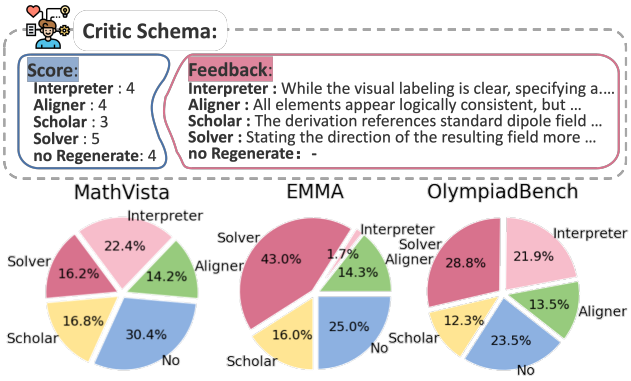

Figure 5: The schema of the *Critic* agent, as well as the feedback and backtracking situations of the *Critic* agent across different datasets.

Variation	MECO	MZCE	MZCO	PECO	PZCE	Avg.
MAPS	46.00	30.33	32.14	28.51	28.33	31.14
$w/o_{\text{Interpreter}}$	25.33	16.67	10.71	21.05	11.62	15.05
Δ	(-20.67)	(-13.66)	(-21.43)	(-7.46)	(-16.71)	(-16.09)
w/o_{Aligner}	28.00	17.67	16.07	20.83	19.00	20.28
Δ	(-18.00)	(-12.66)	(-16.07)	(-7.68)	(-9.33)	(-10.86)
w/o_{Scholar}	28.00	16.33	30.36	19.96	16.33	19.65
Δ	(-18.00)	(-14.00)	(-1.78)	(-8.55)	(-12.00)	(-11.49)
w/o_{Critic}	34.67	21.67	30.36	23.03	21.67	24.09
Δ	(-11.33)	(-8.66)	(-2.42)	(-5.48)	(-6.66)	(-7.05)

Table 2: Performance under different ablation settings are analyzed. We perform ablation experiments on the solving module $w/o_{\text{Interpreter}}$, w/o_{Aligner}, w/o_{Scholar} or w/o_{Critic} modules to evaluate the impact of removing these components.

provement in Table 1. This reflects strong baseline reasoning. In contrast, EMMA and OlympiadBench show the highest feedback for the *Solver*, especially in the interpretation, alignment, and integration steps. These are the most complex and error-prone stages. Other agents receive comparable and lower feedback, indicating better relative performance in their sub-tasks.

Supplementary Analysis

Ablation Studies

Ablating the *Interpreter* results in the greatest loss of performance. We conduct ablation experiments on the OlympiadBench dataset to evaluate the impact of each module on the overall performance. Table 2 presents the effects of removing the *Interpreter*, *Aligner*, *Scholar*, and *Critic* modules from the MAPS framework. The results show that removing the *Interpreter* agent causes the largest performance degradation, at 16.09%. This is because, in complex problem reasoning, diagrams contain a wealth of valuable information, which serves as an important supplement to the text. Understanding diagrams plays a crucial role in problem-solving.

The removal of the *Critic* agent causes the smallest performance loss. It results in only a 7.05% decrease, underscoring its role in providing feedback and corrections. While this mechanism allows MAPS to backtrack and refine its reasoning, its impact is less than that of other agents. Removing the *Scholar* agent results in 11.49% performance drops, highlighting the importance of searching and integrating domain-specific knowledge. Finally, the removal of the *Aligner* agent causes a 10.86% drop, indicating that while diagram and context alignment is valuable, its effect is smaller compared to other components.

Base Model Generalization

MAPS improves performance across diverse base models. We conduct experiments to verify whether our MAPS framework demonstrates robust generalization across various base LLMs. The results confirm MAPS's robustness and transferability, highlighting its adaptability and consistent performance across different foundation models. To further validate its generalization, we evaluate both Qwen2.5-VL-72B and Gemini 2.0 Flash, showing that MAPS performs well across models of varying scales and capabilities. Figure 6 presents results for three sets of base models. In each set, we compare MLLMs and MAPS on mathematical, physical, and chemical sub-tasks. MAPS consistently outperforms the base models. For example, MAPS_{Qwen} improves Qwen2.5-VL-72B by 12.4% in physics, while MAPS_{Gemini} improves Gemini by 4.2%. Similar gains are observed in math and chemistry, demonstrating MAPS's effectiveness on both open-source and closed-source MLLMs.

Time Efficiency

Simpler formats and lower difficulty yield faster solving times. Solving time efficiency varies by question type, answer type, category, and difficulty, with multiple-choice and integer-type questions being the fastest, while higher difficulties and complex formats require more time. Figure 7 illustrates the solving time efficiency across various dimensions—question types, answer formats, subject categories, and difficulty levels—with all times normalized to a *100s* benchmark.

Predefined structure and conceptual simplicity reduce reasoning time. Multiple-choice questions are solved more quickly thanks to predefined answer options that limit the need for extensive reasoning or exploration. Integer-type answers also show high efficiency, often tied to simpler arithmetic or structured formats requiring minimal inference. General category questions are faster on average, likely due to lower conceptual and reasoning complexity compared to domain-specific tasks. In contrast, open-ended questions demand deeper analysis and justification, leading to longer solving times. Finally, solving efficiency declines with increased difficulty: as question complexity rises, so does the required reasoning time, reflecting greater cognitive and computational demands.

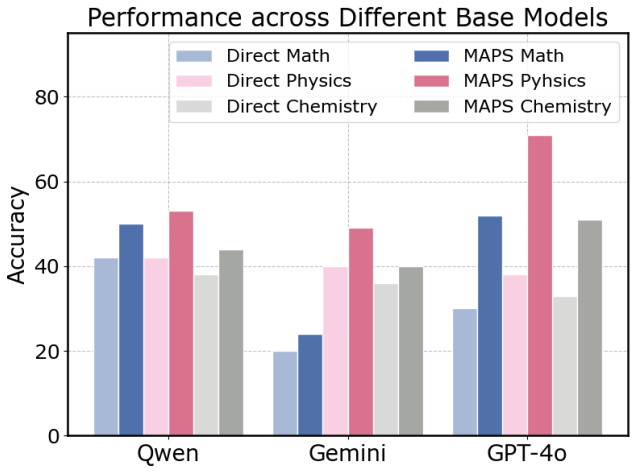

Figure 6: Performance Comparison of MAPS on Math, Physics, and Chemistry Subtasks in the EMMA Dataset with GPT-4o, Gemini, and Qwen2.5-VL-72B as Bases.

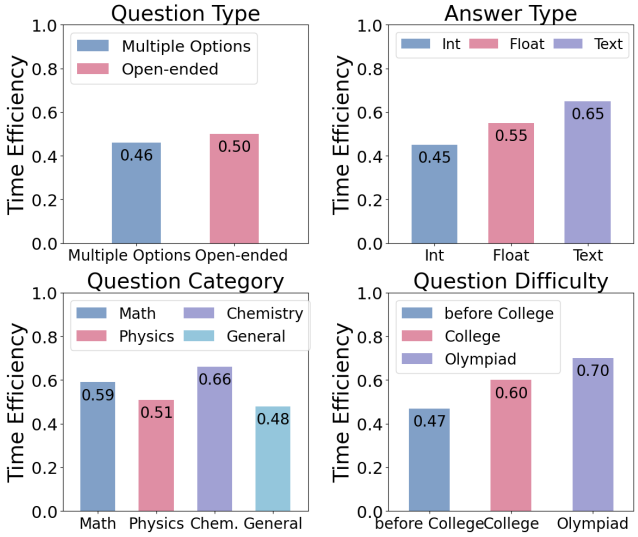

Figure 7: An analysis of the solving time efficiency across different question types, answer types, question categories, and question difficulties.

Related Works

The related work is structured into two main aspects: Firstly, an introduction to complex problem reasoning; Secondly, an exploration of multi-agent techniques.

Complex Problem Reasoning. The research of complex problem reasoning spans across multiple fields, including mathematics, physics, and chemistry, with each area focusing on enhancing problem-solving abilities. In mathematics, studies (Didolkar et al. 2024; Fitriana and Waswa 2024; Tong et al. 2025) explore various methods to improve mathematical problem-solving, such as algorithm optimization, educational strategies, and the use of artificial intelligence. These approaches aim to boost the efficiency, accuracy, and depth of mathematical reasoning. In the field of physics, the papers (Mustofa, Bilad, and Grendis 2024; Kapuriya et al. 2024; Anand et al. 2024; Wu et al. 2024) emphasize the integration of different information types, such as images and text, through multimodal learning to enhance the efficiency and precision of problem solving. In chemistry, three articles (Alasadi and Baiz 2024; Kiernan, Manches, and Seery 2024; Li et al. 2024b; Lin et al. 2025; Dang et al. 2025) investigate the role of multimodal learning in solving chemical problems. By combining diverse information sources, including images and text, and employing techniques such as generative models and molecular geometry reasoning, they aim to improve both the efficiency and accuracy of solving chemistry problems.

Multi-Agent. Multi-agent systems, built on LLMs, consist of multiple AI agents that specialize in specific tasks, working together to solve complex problems (Richards 2023; Yang, Yue, and He 2023; Wu and et al 2023; Sun et al. 2023; Zhang et al. 2026, 2025; Yan et al. 2025; Liu et al. 2025). When presented with a problem, these agents decompose it into smaller, manageable subtasks and utilize various tools, such as internet data retrieval, to solve them through iterative steps. Several studies (Poldrack, Lu, and Beguš 2023; Wang et al. 2024c; Xi et al. 2025; Ni and Gao 2021) have employed multi-agent systems to tackle challenges like problem identification, code writing and debugging, data visualization, and providing interactive feedback to human users. In their work, Ni and Buehler (2024) highlights the potential of AI-driven multi-agent teams in solving mechanical problems autonomously, demonstrating an enhanced capability for understanding, formulating, and validating engineering solutions through self-correction and collaborative refinement. Inspired by the research, we developed the MAPS method, which leverages multi-agent collaborative learning and stepwise problem-solving to provide innovative solutions for complex problem reasoning. By combining the strengths of AI agents, complex problems can be broken down into subtasks and solved step by step through collaboration, improving efficiency and accuracy.

Conclusion

This study presents MAPS, a multi-agent framework grounded in the Big Five Personality Theory and guided by Socratic principles, designed to address the challenges of multimodal comprehensive reasoning and enhance reflective capabilities. The framework involves five agents, each specializing in distinct aspects of problem-solving. To address the first challenge, a four-agent strategy is proposed, where each agent focuses on specific stages of the reasoning process. Additionally, the *Critic* agent addresses the second challenge through Socratic reflection and critical feedback. Extensive experiments on the EMMA, OlympiadBench, and MathVista datasets validate MAPS's effectiveness in overcoming these issues and enhancing performance across various reasoning tasks. Meanwhile, we perform additional analytical experiments to assess the model's advancement as well as its generalization.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). The work is also supported by the National Natural Science Foundation of China (No. 62137002, 62277042, 62293553, 62450005, 62437002, 62477036, 62477037, 62176209, 62192781, 62306229), the "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, the Shaanxi Provincial Social Science Foundation Project (No. 2024P041), the Natural Science Basic Research Program of Shaanxi (No. 2023-JC-YB-593), and the Youth Innovation Team of Shaanxi Universities "Multi-modal Data Mining and Fusion".

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*.

Alasadi, E. A.; and Baiz, C. R. 2024. Multimodal generative artificial intelligence tackles visual problems in chemistry. *Journal of Chemical Education*, 101(7): 2716–2729.

Almagor, M.; Tellegen, A.; and Waller, N. G. 1995. The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. *Journal of personality and social psychology*, 69(2): 300.

Anand, A.; Kapuriya, J.; Singh, A.; Saraf, J.; Lal, N.; Verma, A.; Gupta, R.; and Shah, R. 2024. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 53–64. Springer.

Benet, V.; and Waller, N. G. 1995. The Big Seven factor model of personality description: Evidence for its cross-cultural generality in a Spanish sample. *Journal of Personality and Social Psychology*, 69(4): 701.

Bhattacharya, M.; Pal, S.; Chatterjee, S.; Lee, S.-S.; and Chakraborty, C. 2024. Large language model to multimodal large language model: A journey to shape the biological macromolecules to biological sciences and medicine. *Molecular Therapy-Nucleic Acids*, 35(3).

Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; and Cucchiara, R. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Dang, Z.; Luo, M.; Wang, J.; Jia, C.; Han, H.; Wan, H.; Dai, G.; Chang, X.; and Wang, J. 2025. Disentangled noisy correspondence learning. *IEEE Transactions on Image Processing*.

Didolkar, A.; Goyal, A.; Ke, N. R.; Guo, S.; Valko, M.; Lillicrap, T.; Rezende, D.; Bengio, Y.; Mozer, M.; and Arora, S. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *arXiv preprint arXiv:2405.12205*.

Elder, L.; and Paul, R. 1998. The role of Socratic questioning in thinking, teaching, and learning. *The Clearing House*, 71(5): 297–301.

Fitriana, H.; and Waswa, A. N. 2024. The influence of a realistic mathematics education approach on students' mathematical problem solving ability. *Interval: Indonesian Journal of Mathematical Education*, 2(1): 29–35.

Fu, D.; Guo, R.; Khalighinejad, G.; Liu, O.; Dhingra, B.; Yogatama, D.; Jia, R.; and Neiswanger, W. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*.

Gao, T.; Chen, P.; Zhang, M.; Fu, C.; Shen, Y.; Zhang, Y.; Zhang, S.; Zheng, X.; Sun, X.; Cao, L.; et al. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9096–9105.

Hao, Y.; Gu, J.; Wang, H. W.; Li, L.; Yang, Z.; Wang, L.; and Cheng, Y. 2025. Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark. *arXiv preprint arXiv:2501.05444*.

Hardiansyah, F.; Armadi, A.; AR, M. M.; and Wardi, M. 2024. Analysis of field dependent and field independent cognitive styles in solving science problems in elementary schools. *Jurnal Penelitian Pendidikan IPA*, 10(3): 1159–1166.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z. L.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Kaesberg, L. B.; Becker, J.; Wahle, J. P.; Ruas, T.; and Gipp, B. 2025. Voting or consensus? Decision-making in multi-agent debate. *arXiv preprint arXiv:2502.19130*.

Kapuriya, J.; Kirtani, C.; Singh, A.; Saraf, J.; Lal, N.; Kumar, J.; Shivam, A. R.; Verma, A.; Anand, A.; and Shah, R. R. 2024. Mm-phyrlhf: Reinforcement learning framework for multimodal physics question-answering. *arXiv preprint arXiv:2404.12926*.

Kiernan, N. A.; Manches, A.; and Seery, M. K. 2024. Resources for reasoning of chemistry concepts: multimodal molecular geometry. *Chemistry Education Research and Practice*, 25(2): 524–543.

Landau, R. H.; Páez, M. J.; and Bordeianu, C. C. 2024. *Computational physics: Problem solving with Python*. John Wiley & Sons.

Li, J.; Lu, W.; Fei, H.; Luo, M.; Dai, M.; Xia, M.; Jin, Y.; Gan, Z.; Qi, D.; Fu, C.; et al. 2024a. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Li, J.; Zhang, D.; Wang, X.; Hao, Z.; Lei, J.; Tan, Q.; Zhou, C.; Liu, W.; Yang, Y.; Xiong, X.; et al. 2024b. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*.

Li, L.; Wang, Y.; Xu, R.; Wang, P.; Feng, X.; Kong, L.; and Liu, Q. 2024c. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Lin, Q.; Zhu, Y.; Pu, B.; Huang, L.; Luo, H.; Ma, J.; Peng, Z.; Zhao, T.; Xu, F.; Zhang, J.; et al. 2025. A Foundation Model for Chest X-ray Interpretation with Grounded Reasoning via Online Reinforcement Learning. *arXiv preprint arXiv:2509.03906*.

Liu, W.; Luo, H.; Lin, X.; Liu, H.; Shen, T.; Wang, J.; Mao, R.; and Cambria, E. 2025. Prompt-R1: Collaborative Automatic Prompting Framework via End-to-end Reinforcement Learning. arXiv:2511.01016.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Mustofa, H. A.; Bilad, M. R.; and Grendis, N. W. B. 2024. Utilizing AI for physics problem solving: a literature review and ChatGPT experience. *Lensa: Jurnal Kependidikan Fisika*, 12(1): 78–97.

Ni, B.; and Buehler, M. J. 2024. MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. *Extreme Mechanics Letters*, 67: 102131.

Ni, B.; and Gao, H. 2021. A deep learning approach to the inverse problem of modulus identification in elasticity. *Mrs Bulletin*, 46: 19–25.

Paul, R.; and Elder, L. 2019. *The thinker's guide to Socratic questioning*. Rowman & Littlefield.

Poldrack, R. A.; Lu, T.; and Beguš, G. 2023. AI-assisted coding: Experiments with GPT-4. *arXiv preprint arXiv:2304.13187*.

Qiu, J.; Yuan, W.; and Lam, K. 2024. The application of multimodal large language models in medicine. *The Lancet Regional Health–Western Pacific*, 45.

Richards, T. B. 2023. Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous.

Simms, L. J. 2007. The Big Seven model of personality and its relevance to personality pathology. *Journal of Personality*, 75(1): 65–94.

Sun, Q.; Yin, Z.; Li, X.; Wu, Z.; Qiu, X.; and Kong, L. 2023. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *arXiv preprint arXiv:2310.00280*.

Tong, Y.; Zhang, X.; Wang, R.; Wu, R.; and He, J. 2025. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37: 7821–7846.

Wang, K. D.; Burkholder, E.; Wieman, C.; Salehi, S.; and Haber, N. 2024a. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. In *Frontiers in Education*, volume 8, 1330486. Frontiers Media SA.

Wang, L.; Hu, Y.; He, J.; Xu, X.; Liu, N.; Liu, H.; and Shen, H. T. 2024b. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17, 19162–19170.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024c. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wu, Q.; and et al. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Wu, W.; Zhang, L.; Liu, J.; Tang, X.; Wang, Y.; Wang, S.; and Wang, Q. 2024. E-gps: Explainable geometry problem solving via top-down solver and bottom-up generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13828–13837.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2): 121101.

Xu, F.; Lin, Q.; Han, J.; Zhao, T.; Liu, J.; and Cambria, E. 2025. Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond. *Transactions on Knowledge and Data Engineering*, 37: 1620–1634.

Yan, H.; Xu, F.; Xu, R.; Li, Y.; Zhang, J.; Luo, H.; Wu, X.; Tuan, L. A.; Zhao, H.; Lin, Q.; et al. 2025. Mur: Momentum uncertainty guided reasoning for large language models. *arXiv preprint arXiv:2507.14958*.

Yang, H.; Yue, S.; and He, Y. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; and Yu, D. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Zhang, J.; Wang, Z.; Zhu, H.; Liu, J.; Lin, Q.; and Cambria, E. 2026. MARS: A Multi-Agent Framework Incorporating Socratic Guidance for Automated Prompt Optimization. In *Proceedings of AAAI*.

Zhang, J.; Wei, B.; Qi, S.; Liu, J.; Lin, Q.; et al. 2025. GKG-LLM: A Unified Framework for Generalized Knowledge Graph Construction. *arXiv preprint arXiv:2503.11227*.