FlowSearch: Advancing Deep Research with Dynamic Structured Knowledge Flow

Yusong Hu* **Runmin Ma*** **Yue Fan*** **Jinxin Shi** **Zongsheng Cao**
Yuhao Zhou **Jiakang Yuan** **Shuaiyu Zhang** **Shiyang Feng** **Xiangchao Yan**
Shufei Zhang **Wenlong Zhang** **Lei Bai†** **Bo Zhang†**
Shanghai Artificial Intelligence Laboratory
{zhangbo,bailei}@pjlab.org.cn

https://github.com/InternScience/InternAgent

Abstract

Deep research is an inherently challenging task that demands both breadth and depth of thinking. It involves navigating diverse knowledge spaces and reasoning over complex, multi-step dependencies, which presents substantial challenges for agentic systems. To address this, we propose FlowSearch, a multi-agent framework that actively constructs and evolves a dynamic structured knowledge flow to drive subtask execution and reasoning. FlowSearch is capable of strategically planning and expanding the knowledge flow to enable parallel exploration and hierarchical task decomposition, while also adjusting the knowledge flow in real time based on feedback from intermediate reasoning outcomes and insights. FlowSearch achieves competitive performance on both general and scientific benchmarks, including GAIA, HLE, GPQA and TRQA, demonstrating its effectiveness in multi-disciplinary research scenarios and its potential to advance scientific discovery. The code will be available.

1 Introduction

The general capabilities of Large Language Models (LLMs) enable agent systems for diverse tasks, supporting scientific research and discovery (Schick et al., 2023; Fan et al., 2024; Team et al., 2025a; Zhang et al., 2025). However, effectively utilizing these capabilities in open-ended research requires iterative hypothesis formulation, strategic information acquisition, and multi-step reasoning in dynamic and uncertain knowledge spaces. These challenges have inspired the development of Deep Research (DR), which combines LLMs with expert-designed knowledge to go beyond basic reasoning or simple information retrieval. Such systems are essential for unlocking the potential of LLMs in enabling scientific discovery on various domains.

Existing deep research systems (Hu et al., 2025; Team, 2025) are primarily inspired by either individual or collaborative research paradigms. **(1) Single-agent paradigm:** (Wu et al., 2025; Tao et al., 2025; Li et al., 2025b; Yao et al., 2023b) a single LLM centrally manages the research workflow using a long context window to accumulate and reason over information. While this mirrors individual researchers, it is prone to tunnel vision, overcommitting to early hypotheses and limiting exploratory breadth. **(2) Multi-agent paradigm:** (Hu et al., 2025; man, 2025; Team et al., 2025a) research is scaled via explicit planning and role specialization. However, serial plan execution requires maintaining context across multiple planning steps and agents, forcing aggressive context truncation or retrieval filtering. As a result, intermediate reasoning chains are often fragmented, leading to shallow, step-wise reasoning and a loss of global reasoning depth (Huang et al.). Overall, both paradigms trade off exploratory breadth against reasoning depth, motivating the need for more adaptive and context-aware research agents.

In this work, we propose a novel Dynamic Structured Knowledge Flow to enable structured and efficient knowledge propagation throughout scientific discovery activities, and instantiate it in FlowSearch, a multi-agent system grounded in this design. Unlike conventional sequential planning, the proposed Structured Knowledge Flow models scientific discovery as **an evolving graph** (as illustrated in Fig. 3) of interdependent knowledge units, where exploratory breadth and reasoning depth are jointly supported through incremental expansion and structural revision. Within this flow, knowledge is decomposed, integrated, and summarized at different granularity levels, enabling deep local reasoning while preserving global coherence.

FlowSearch instantiates this abstraction as a multi-agent system by operationalizing the dy-

*Equal Contribution
†Corresponding Authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21212–21245

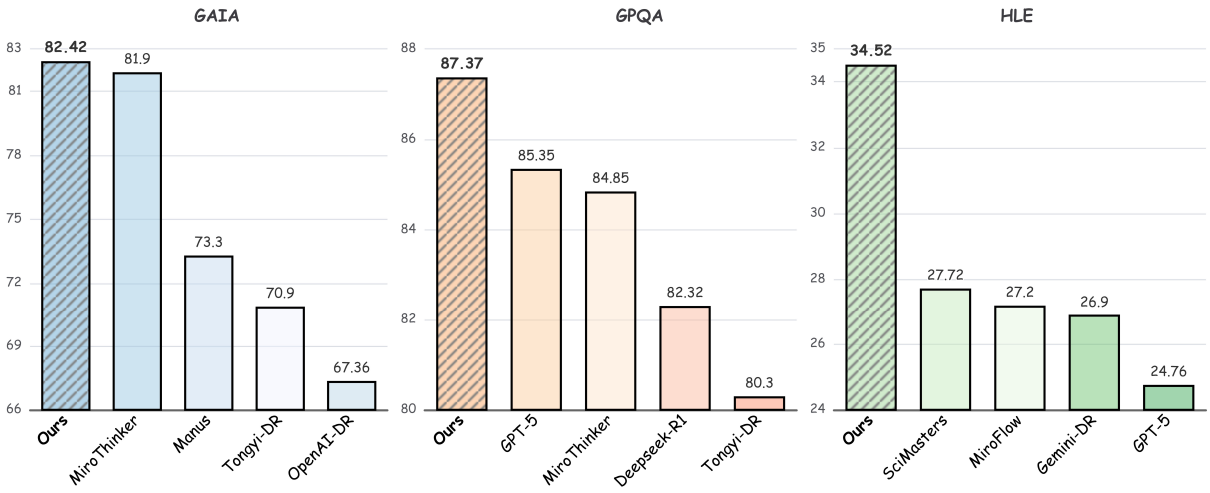

Figure 1: FlowSearch (Ours) achieves leading performance on the GAIA, GPQA, and HLE benchmarks, outperforming competitive agentic frameworks (OpenAI-DR, MiroFlow, Tongyi-DR and Manus) as well as LLM-based approaches (GPT-5, DeepSeek-R1, MiroThinker).

namic structured knowledge flow throughout the research process. It starts with a flow planner that initializes a graph-structured knowledge flow, where nodes represent subproblems or key concepts and edges capture their knowledge dependencies. As scientific discovery progresses, the knowledge flow is continuously expanded and dynamically revised in response to intermediate findings. Each node orchestrates the execution of corresponding subtasks while supporting recursive decomposition, integration of upstream knowledge, and local summarization, thereby enabling structured and efficient knowledge propagation with effective context management across complex, multi-step scientific problem solving.

We conduct experiments on several challenging benchmarks, including GAIA (Mialon et al., 2023), which evaluates the general problem-solving abilities of AI assistants, as well as three scientific-question-answering benchmarks HLE (Phan et al., 2025), GPQA (Rein et al., 2024) and TRQA (Zhang et al., 2025). As shown in Fig. 1, our approach achieves state-of-the-art results on GAIA, HLE, TRQA and GPQA. These findings highlight the strong problem-solving capability of FlowSearch, enabled by the integration of graph-driven planning. In summary, our main contribution can be described as follows:

- Unlike conventional sequential frameworks in deep research agents, we introduce a novel dynamic structured knowledge flow to achieve a trade-off between exploratory breadth and reasoning depth.
- We develop FlowSearch, a multi-agent system built upon the dynamic structured knowledge flow, capable of generating structured plans and dynamically refining them during execution phase to enhance performance.
- We evaluate FlowSearch on the general AI assistant benchmark GAIA and the multi-disciplinary scientific benchmarks HLE, GPQA and TRQA, demonstrating state-of-the-art results.

2 Related Work

2.1 Agentic Systems

Agentic systems with LLM have evolved from static prompting to perception–action loops, enabling systems to plan (Wang et al., 2023b), act (Yao et al., 2023b), and learn using external tools. Foundational approaches, such as interleaved reasoning–acting frameworks (Yao et al., 2023b) and tree search planning (Yao et al., 2023a), improve reliability in multi-step tasks, while reflective self-revision mechanisms (Shinn et al., 2023) and external memory (Wang et al., 2023a) enhance long-horizon consistency. Recent efforts like OpenHands (Wang et al., 2025) further expands action spaces and mitigate hallucinations, and evaluates on more realistic interactive benchmarks including AgentBoard (Ma et al., 2024), StuLife (Cai et al., 2025), and SWE-bench Verified (bench Team, 2024). Multi-agent orchestration involves role-specialized collaboration and negotiation (Zhuge et al., 2024), replacing single-agent end-to-end optimization with a modular and scalable approach.

Despite these advances, most general-purpose

agents target short to medium horizon tasks and interactive environments. Scientific research (OpenAI, 2025b), however, requires handling long-horizon workflows, integrating diverse knowledge, and adapting strategies dynamically. This motivates the development of research-oriented agents, which focus on structured, adaptive, and knowledge-driven scientific inquiry.

2.2 Deep Research Agents

Recent advances in DR agents extend LLMs from retrieval-augmented generation to dynamic, tool-driven research workflows. Early systems such as WebGPT (Nakano et al., 2021) and Toolformer (Schick et al., 2023) explored web and API integration, while industrial solutions *e.g.*, OpenAI DR (OpenAI, 2025b), Gemini DR (Google, 2024), Grok DR (xAI, 2025), Perplexity DR (Perplexity, 2025), incorporate adaptive planning, iterative retrieval, and multimodal reasoning. Recently, single-agent designs (*e.g.*, Search-o1 (Li et al., 2025a), WebDancer (Wu et al., 2025), Tongyi DeepResearcher (Qiao et al., 2025)) enable end-to-end optimization, while multi-agent architectures (*e.g.*, AI Scientist (Lu et al., 2024), Agent Laboratory (Schmidgall et al., 2025), and InternAgent (Team et al., 2025a)) offer modularity and scalability—crucial for complex research.

Recent studies, *e.g.*, GeAR (Shen et al., 2024), PANGU DeepDiver (Shi et al., 2025) also show the benefit of explicit structures and self-evolving mechanisms for multi-hop reasoning. However, the existing DR agents still suffer from sequential bottlenecks and limited hierarchical decomposition, motivating frameworks like our FlowSearch that integrate multi-agent coordination with dynamic structured knowledge flow.

3 FlowSearch

FlowSearch is built upon a **Dynamic Structured Knowledge Flow** that enables structured and adaptive scientific research, which allows the system to follow a clear and coherent research trajectory, while dynamically revising local reasoning paths based on intermediate findings to ensure the correctness of ongoing investigation. As illustrated in Fig. 2, FlowSearch comprises three core components: **Knowledge Flow Planner**, which constructs high-quality knowledge flows tailored to the research objective; **Knowledge Collector**, which executes subtasks and enriches each node with relevant contextual information; and **Knowledge Flow Refiner**, which monitors progress and dynamically adjusts the flow based on intermediate outcomes and newly acquired knowledge. We begin by formalizing the concept of the structured Dynamic Structured Knowledge Flow, followed by detailed descriptions of Knowledge Flow Planner, Knowledge Collector, and Knowledge Flow Refiner.

3.1 Structured Knowledge Flow

Structured Knowledge Flow provides principled guidance for systematically organizing information to improve both the systematicity and effectiveness of deep research. Moreover, its graph-structured formulation enables flexible revision of local research objectives and knowledge dependencies when necessary.

A common practice in deep research agents is to address a user query q by assembling a strictly linear pipeline $L(q) = [a_1, a_2, \ldots, a_n]$ and executing it in order $a_1 \to a_2 \to \cdots \to a_n$. The precedence relations are implicitly encoded by positional order, *i.e.*, $a_i \prec a_j \iff i < j$. Despite its procedural simplicity and ease of implementation, such a linear formalism fails to capture the inherently complex and non-linear dependencies of real-world research processes.

To better capture the complex structure of deep research reasoning processes, we adopt a directed acyclic graph $G = (V, E)$ to explicitly model both task dependencies and knowledge flow. Each node $v_i \in V$ is a typed subtask node $v_i = (t_i, d_i, s_i, c_i)$, where $t_i \in \{search, solve, answer\}$ is the task type, d_i is the task desciption, s_i is the execution state of the node and c_i is the resulting knwoledge context of the node if successfully executed. Each directed edge $e_{ij} = (v_i, v_j, r_{ij}) \in E$ specifies how the output of v_i conditions or constrains v_j using the relation $r_{ij} \in R$, where R is the set of relation types. This flow makes precedence among the nodes and supports parallel execution on independent branches, yielding a more expressive and verifiable substrate for Deep Research.

As an illustration, the following example describes a minimal graph in natural language form:

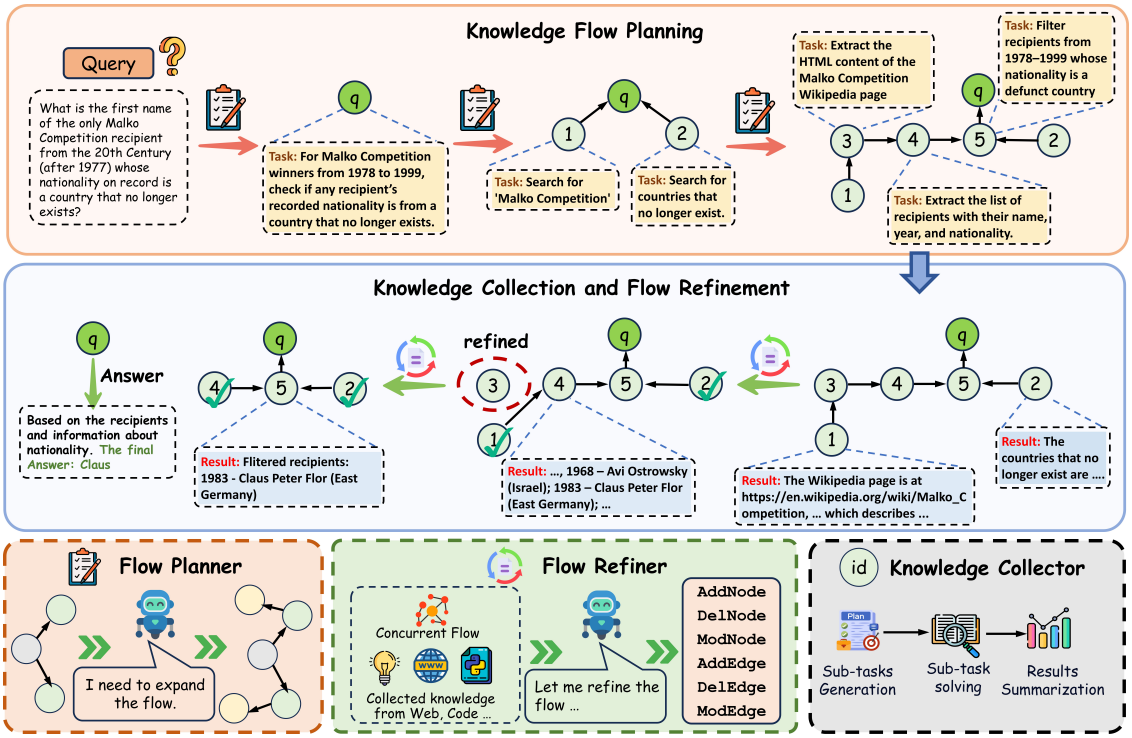

Figure 2: Overview of FlowSearch. **Top part** illustrates the Knowledge Flow Planning process, where the Knowledge Flow Planner incrementally expands the structured knowledge flow. **Middle part** depicts the iterative process of Knowledge Collection and Flow Refinement, where nodes are executed by the Knowledge Collector and the flow is dynamically adjusted by the Knowledge Flow Refiner based on newly acquired knowledge. **Lower part** shows the three key components of FlowSearch—Flow Planner (left), Flow Refiner (center), and Knowledge Collector (right)—and their collaborative role in enabling systematic, adaptive, and efficient deep research.

```
{
 "nodes": [
  {"node_id": "n1", "task_type": "answer", "content": "<query>"},
  {"node_id": "n2", "task_type": "solve", "content": "<subtask>"},
  {"node_id": "n3", "task_type": "search", "content": "<subtask>"},
 ],
 "edges": [
  {"from": "n2", "to": "n1", "relationship": "solve subtask"},
  {"from": "n3", "to": "n1", "relationship": "provide information"},
 ]
}
```

By formalizing the research process in this manner, the agent is enabled not only to generate execution plans, but also to reason over the structural dependencies among subtasks, ensuring coherence and systematicity throughout the deep research workflow. Moreover, the graph-structured formulation allows multiple non-dependent nodes to be executed in parallel, and enables flexible revision of local research components without interfering with other ongoing processes, including the insertion of new nodes or the modification of suboptimal planning decisions when errors are identified, thereby enhancing the robustness of the overall system. These mechanisms will be further elaborated in the subsequent section on the construction of FlowSearch.

3.2 Knowledge Flow Planner

A high-quality Knowledge Flow is essential for the effective execution of complex research tasks. Rather than constructing the entire structure in a single step, which can lead to instability and reduced control, we employ a Knowledge Flow Planner process that incrementally initializes the flow.

Let $G_t^{init} = \{V_t^{init}, E_t^{init}\}$ be the flow in the t-th initialization iteration. Specifically, $G_0^{init} = \{\{v_{query}\}, \emptyset\}$ only contains the query node at the beginning. At each iteration t, an LLM planner examines the nodes in the current flow G_t^{init} to identify those requiring further decomposition or additional context. For each node requires decomposition, the planner generates a set of successor nodes representing sub-questions, reasoning steps, or supporting evidence for it. The corresponding dependency edges are added to the flow to maintain structural coherence and preserve logical consistency, which can be formulated as follows:

$$G_{t+1}^{init} = \{V_{t+1}^{init}, E_{t+1}^{init}\} = f^{expand}(G_t^{init}), \quad (1)$$

where $f^{expand}(\cdot)$ is a prompted LLM planner, $V_{t+1}^{init} = V_t^{init} \cup V_t^{add}$ contains newly added nodes

and $E_{t+1}^{init} = E_t^{init} \cup E_t^{add}$ contains newly introduced edges (dependencies) connecting nodes in V_t^{add}. This iterative expansion progressively extends the boundaries of the research and deepens the level of exploration within the knowledge flow. The process continues until $f^{expand}(\cdot)$ yields no additional nodes. Upon completion of the expansion phase, an initial flow $G_0 = G_T^{init}$ is instantiated to support subsequent knowledge collector and flow refiner, where T is the iteration steps in the flow expansion stage.

After the initial planning of the knowledge flow, FlowSearch then enters another iterative loop of Knowledge Collector and Flow Refiner. This iteration continues until the original user query is successfully resolved. Knowledge Collector and Flow Refiner are described as follows.

3.3 Knowledge Collector

The Knowledge Collector aims at identifying the outermost executable nodes in the flow—those whose dependencies have all been resolved—and assigns each to an executor agent for processing. These agents, implemented as large language models equipped with tools, decompose the subtask into a sequential execution trajectory, iteratively reasoning and retrieving information to resolve the node. Available tools include web browsing, file downloading, and visual question answering, etc. A complete list of supported tools is provided in the Appendix A.

After the execution of node v_i, its execution state s_i (either success or failure) is updated. If the execution succeeds, the resulting knowledge—either retrieved or derived through reasoning—is distilled into a summarized knowledge context c_i, which serves as input for the subsequent execution of the nodes that depend on it. Formally, given $G_t = (V_t, E_t)$ in the t-th Knowledge Collector and Flow Refiner iteration, the execution of node v_i can be described as:

$$s_i, c_i = f^{exec}(t_i, d_i | \{c_j | (v_j \to v_i) \in E_t\}), \quad (2)$$

where t_i and d_i are the task type and task description of node v_i, $f^{exec}(\cdot)$ is the LLM executor with tools depending on the context knowledge $\{c_j | (v_j \to v_i) \in E_t\}$.After the parallel execution of all the outermost executable nodes is completed, a flow refinement will be conducted based on the newly obtained knowledge, which will be described in Sec. 3.4.

3.4 Knowledge Flow Refiner

After completing the execution of nodes and updating the corresponding knowledge in each iteration, FlowSearch activates the Knowledge Flow Refiner to check the knowledge flow for potential factual inconsistencies. Leveraging the knowledge collected by the Knowledge Collector, Knowledge Flow Refiner analyzes the current flow and identify potential structural adjustments, including the addition, removal, or modification of tasks and dependencies. The goal of Knowledge Flow Refiner is to advance the research task in a reflective way and enhance execution efficiency.

The Knowledge Flow Refineris prompted to utilize a set of predefined graph transformation operations to modify nodes and edges in the flow based on the knowledge context and execution states of the existing nodes. These operations include:

- **Add Node** (`AddNode`): Introduce new nodes to capture missing sub-questions, intermediate reasoning steps, or evidence that were not anticipated in the initial flow.
- **Delete Node** (`DelNode`): Remove nodes that are redundant, irrelevant, or no longer necessary given the updated knowledge.
- **Modify Node** (`ModNode`): Modify the attributes of current nodes, especially the content of the sub-task.
- **Add Edge** (`AddEdge`): Create new dependency edges to reflect newly discovered relationships between nodes.
- **Delete Edge** (`DelEdge`): Remove edges that represent incorrect, obsolete, or redundant dependencies, ensuring a more reasonable graph structure.
- **Modify Edge** (`ModEdge`): Modify existing edges to correct dependency directions or improve the structure for more efficient execution.

Formally, $G_{t+1} = f^{refine}(\{V_t, E_t\})$, where f^{refine} is an LLM that generates a sequence of graph transformation operations $\mathcal{O} = \{o_1, o_2, \ldots, o_m\}$ and applies them on $G_t = \{V_t, E_t\}$ to obtain the updated flow G_{t+1}. Through ongoing adjustments, FlowSearch achieves coherent and goal-directed reasoning.

3.5 Conclusion Generation

At the end of all iterations of knowledge collector and flow refiner, only the initial query node remains

Table 1: Performance comparison on GAIA, GPQA-diamond and HLE benchmarks. The best results are **bolded** and the second best results are underlined. Results not reported in the original papers are denoted as " - ". **Note that** some approaches, such as MiroThinker and TongyiDR, can only report results on the HLE text-only subset as these approaches lack multimodal ability.

		GAIA val				GPQA-diamond				HLE	
Method	**Base Model**	Level 1	Level 2	Level 3	Avg.	Bio	Chem	Phys	Avg.	text only	All
No Agency											
Qwen-3-8B	-	11.32	2.32	0.00	4.85	-	-	-	44.44	-	-
Qwen3-32B	-	13.21	3.49	3.84	6.67	-	-	-	49.49	-	-
Qwen3-235B	-	15.09	3.49	3.84	7.27	-	-	-	47.47	9.18	8.60
Intern-S1	-	28.30	9.30	7.69	15.15	**89.47**	59.49	93.02	78.26	8.90	8.30
Deepseek-R1	-	33.96	13.95	3.84	18.78	63.16	76.34	91.86	82.32	8.60	-
o4-mini	-	28.30	12.79	7.69	16.97	78.95	63.44	94.19	78.28	14.50	14.28
GPT-5	-	-	-	-	-	84.21	76.34	95.35	85.35	25.85	24.76
Close-sourced Agentic Framework											
OpenAI DR	-	74.29	69.06	47.60	67.36	-	-	-	-	-	26.60
Manus	-	86.50	70.10	57.70	73.30	-	-	-	-	-	-
Gemini DR	-	-	-	-	-	-	-	-	-	-	26.90
React Agentic Model											
WebDancer	QwQ-32B	61.5	50.0	25.0	51.5	-	-	-	-	-	-
WebShaper	Qwen2.5-72B	69.2	63.4	16.6	60.1	47.37	52.69	81.40	64.65	-	-
MiroThinker	Qwen2.5-72B	-	-	-	81.9	84.21	75.27	95.35	84.85	**37.7**	-
Tongyi DR	Qwen3-30B-A3B	-	-	-	70.9	78.95	67.74	95.35	80.30	32.9	-
Open-sourced Agentic Framework											
MiroFlow	Claude-3.7	-	-	-	74.50	-	-	-	-	29.50	27.20
OWL	Gemini-2.5-Pro	84.90	68.60	42.30	69.70	57.89	61.29	86.05	71.72	-	-
X-Masters	Deepseek-R1	-	-	-	-	78.95	68.82	94.19	80.81	32.10	27.72
JoyAgent	Claude-4	86.79	77.91	42.31	75.15	78.95	65.59	91.86	77.27	-	-
AWorld	Gemini-2.5-Pro	88.68	77.91	53.85	77.58	73.68	66.67	93.02	78.79	-	-
FlowSearch											
FlowSearch	Qwen3-235B	69.81	60.47	30.77	58.79	63.16	58.06	75.58	66.16	15.04	14.84
FlowSearch	o4-mini	**92.45**	**82.56**	**61.54**	**82.42**	84.21	**79.57**	**96.51**	**87.37**	36.10	**34.52**

unexecuted. If the query is a straightforward scientific question that can be answered directly and simply, the query node will utilize the knowledge from its connected nodes to provide an immediate response. If the query requires generating a detailed scientific report, the final query node will aggregate knowledge from all nodes within the flow, perform a comprehensive reasoning process, and deliver a complete and thorough report. The details are explained in Appendix B.

4 Experiments

To comprehensively assess the capabilities of FlowSearch, we conduct experiments on a diverse set of challenging benchmarks, ranging from general question answering to scientific deep research.

4.1 Experiments Setup

Evaluation Benchmarks. We conduct extensive experiments on four challenging benchmarks:

- **GAIA** (Mialon et al., 2023): a benchmark of real-world questions that require a set of fundamental abilities such as reasoning, multimodality handling, web browsing, and generally tool-use proficiency. Our results are reported on its 165-question validation set.

- **GPQA** (Rein et al., 2024): a benchmark of 448 multiple-choice questions across biology, chemistry, and physics, authored by domain experts to ensure depth and rigor, thereby providing a stringent evaluation of advanced reasoning and scientific knowledge. We use its 198-question GPQA-diamond subset for evaluation.

- **HLE** (Phan et al., 2025): Humanity's Last Exam is a multimodal benchmark consisting of 2,500 questions across mathematics, humanities, and natural sciences. Developed by subject experts, it provides a frontier-level test

Table 2: Ablation study on the impact of structured planning and refinement. We compare the workflow with conventional sequential planner, the flow planner, and the flow refiner. A checkmark (✓) indicates the component is used. Results are reported on GAIA and GPQA.

			GAIA				GPQA			
Sequential Planner	**Flow Planner**	**Refiner**	Level 1	Level 2	Level 3	Avg	Bio	Chem	Phys	Avg
✓	–	–	67.92	55.81	23.07	55.76	57.89	54.84	88.37	71.21
–	✓	–	73.58	63.95	30.77	61.82	57.89	59.14	89.53	73.74
–	✓	✓	**92.45**	**82.56**	**61.54**	**82.42**	**84.21**	**79.57**	**96.51**	**87.37**

of academic competence where current LLMs still perform far below human experts.

- **TRQA** (Zhang et al., 2025): a domain-specific benchmark for therapeutic target discovery. It covers fundamental biology, disease biology, pharmacology, and clinical medicine, providing a systematic evaluation framework for biomedical research agents. We use its 172-question TRQA-lit subset for evaluation.

Competing Methods. To validate the effectiveness of FlowSearch, we compare FlowSearch on GAIA, GPQA, HLE and TRQA against cutting-edge large language models including Qwen3 series model (Yang et al., 2025), Intern-S1 (Bai et al., 2025), Deepseek-R1 (Guo et al., 2025), GPT-o4-mini and GPT-5 (OpenAI, 2025a), and some state-of-the-art deep research agent, including proprietary approaches OpenAI-DR (Deep Research) (OpenAI, 2025b), Gemini-DR (Google, 2024) and Manus (man, 2025), leading react agentic models MiroThinker (Team, 2025), Tongyi-DR (Team et al., 2025b), WebShaper (Tao et al., 2025), WebDancer (Wu et al., 2025), and open-source frameworks MiroFlow (Team, 2025), OWL (Hu et al., 2025) X-Masters (Chai et al., 2025), JoyAgent (Liu et al., 2025), AWorld (Yu et al., 2025) and Origene (Zhang et al., 2025). In the experiments, we utilize GPT-o4-mini to serve as the Knowledge Flow Planner, Knowledge Collector and Knowledge Flow Refiner in our workflow.

4.2 Experiment Results

Table 1 and Fig. 4 present the performance of FlowSearch and its counterparts on GAIA, GPQA, HLE, and TRQA. FlowSearch consistently achieves competitive results across all benchmarks without training an additional base model, validating the effectiveness of its systematic design.

4.2.1 Agentic Benchmark

FlowSearch achieves state-of-the-art performance among agentic systems. On GAIA (Table 1), FlowSearch (o4-mini) outperforms both closed-source Manus (73.30%) and leading open-source agentic models Mirothinker (81.9%) and Tongyi-DR (70.9%), even though they are specifically trained and evaluated only on the GAIA text-only subset. Flowsearch also shows strong robustness on Level 3 questions (61.54%). These results indicate that its iterative workflow combining knowledge planning, collection, and refinement is particularly effective for multi-hop and compositional reasoning. Its clear advantage over systems like OpenAI DR and MiroFlow further underscores the impact of structured and dynamic workflow design.

Agentic systems consistently outperform pure LLMs on complex reasoning tasks. Larger base models like the Qwen series benefit from greater internal knowledge, but remain limited without structured reasoning. Even models fine-tuned for scientific reasoning, such as Intern-S1, lag behind agentic approaches. For instance, FlowSearch with o4-mini achieves a score of 82.42% on GAIA, far surpassing the same model (o4-mini) without agency (16.97%), highlighting that structured task decomposition and flow-based execution are more critical than model size alone.

4.2.2 Multi-Disciplinary Research and Question Answering

Although scientific benchmarks are generally knowledge-intensive and favor LLMs, FlowSearch still achieves competitive performance.

FlowSearch effectively acquires domain-specific knowledge through Knowledge Flow. On the GPQA-diamond benchmark, FlowSearch (o4-mini) achieves 87.37% average accuracy in Biology, Chemistry, and Physics—outperforming GPT-5 and Intern-S1. This underscores the advantage of dynamic retrieval in accessing context-relevant knowledge, which enables more accurate and flexible scientific reasoning than relying solely on static information.

General-purpose tools guided by Knowledge

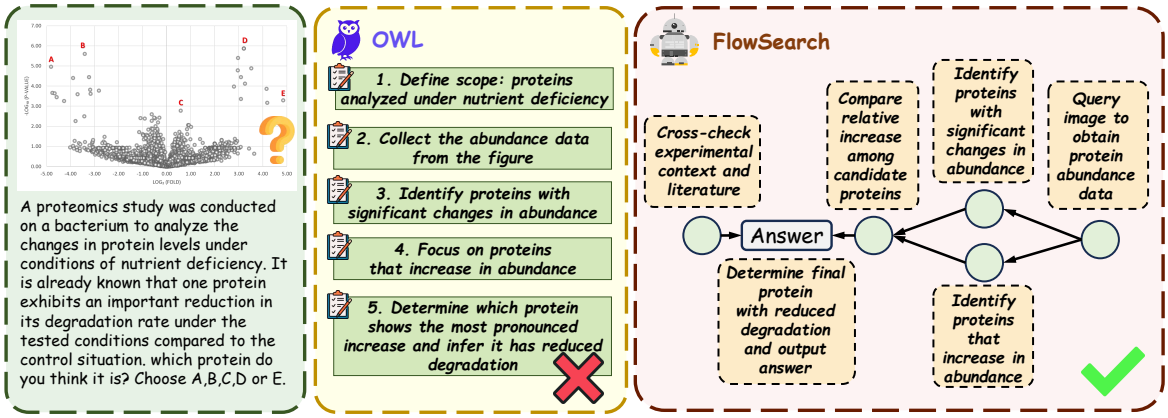

Figure 3: Case study comparing the conventional deep research framework OWL with our FlowSearch.

Flow can outperform specialized systems. On the HLE benchmark, FlowSearch (o4-mini) achieves the highest accuracy among training-free methods at 34.52%, demonstrating performance comparable to leading trained agentic models such as MiroThinker (37.7%) and Tongyi-DR (32.9%) on the text-only subset, as well as closed-source systems including OpenAI DR (26.60%) and Gemini Deep Research (26.90%). On TRQA, FlowSearch reaches 77.9%, outperforming domain-specific agent Origene (60.1%) and scientific multi-modal model Intern-S1 (49.4%). These results show that a well-structured general-purpose agent system can effectively tackle complicated scientific and cross-domain tasks.

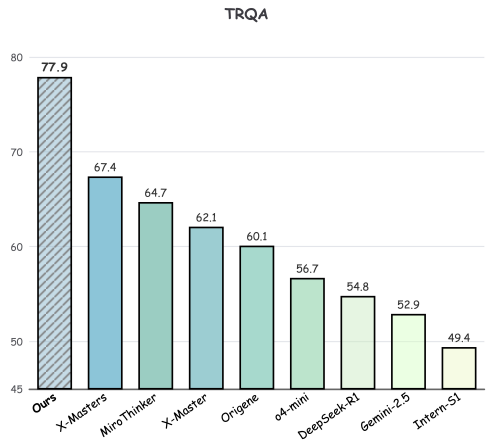

Figure 4: Performance on TRQA. FlowSearch significantly outperforms previous works.

4.3 Analysis and Visualization

Ablation on Key Components. We conduct ablation studies on two critical components of FlowSearch: the Knowledge Flow Planner and the Knowledge Flow Refiner. As shown in Table 2, replacing conventional sequential planner reasoning with our structured Knowledge Flow leads to substantial performance improvements, with gains of 6.06% on GAIA and 2.53% on GPQA. This highlights its effectiveness in capturing complex task dependencies and enhancing problem-solving capabilities. Moreover, incorporating the Flow Refiner yields further notable improvements, indicating that dynamic flow refinement enables more flexible task adaptation and strengthens the agent's overall research competence.

Case Study and Visualization. Fig. 3 illustrates the contrast between our knowledge-flow-based FlowSearch and the conventional sequential planning paradigm, represented by OWL (Hu et al., 2025), in addressing a scientific question. As shown in the figure, OWL decomposes the query into a linear sequence of subtasks—such as understanding, information collection, and identification—that are executed in order. While this pipeline is straightforward, it lacks mechanisms to preserve and integrate intermediate insights, which leads to the dilution of valuable evidence as the chain grows longer.

Compared to other methods, FlowSearch constructs a structured knowledge flow directly from the user query, explicitly modeling dependencies between subtasks. For example, it ensures that a question about an image is asked only after extracting relevant information from the image.Each node within the flow performs its designated operation, summarizes the outcome, and passes structured intermediate results to subsequent steps. This design facilitates the reuse of prior knowledge, minimizes the propagation of irrelevant information, and ensures that critical evidence is preserved throughout the reasoning process.

5 Conclusion

In this work, we have presented FlowSearch, a multi-agent deep research built on a dynamic structured knowledge flow. By explicitly modeling dependencies among subproblems and key concepts, the system enables both deep reasoning within local regions of the knowledge flow and coherent knowledge propagation at a global level. Experimental results highlight the effectiveness of combining structured knowledge flow planning, suggesting that such frameworks offer a promising direction for building autonomous and reflective systems for tackling complex scientific research tasks.

Limitations

Despite the strong empirical performance of FlowSearch, its effectiveness currently depends on access to high-capability proprietary foundation models. In our experiments, substituting these models with existing open-source alternatives leads to a noticeable performance degradation, especially in complex reasoning and long-horizon planning scenarios.

In terms of efficiency, although FlowSearch benefits from parallel execution and iterative reflection, which improve sample efficiency compared to sequential baselines, the computational and monetary costs remain non-trivial when deployed at scale. Large-scale applications that require extensive tool use, repeated planning, and verification may still incur substantial overhead, highlighting the need for further optimization and more cost-effective model alternatives.

Ethics Statement

This work studies FlowSearch, a deep research framework for multi-disciplinary research. All benchmarks used in our experiments are publicly available and widely adopted in prior work. We conduct evaluations under consistent protocols, with identical tool access and comparable inference budgets across methods to ensure fair and transparent comparison. Our experiments strictly adhere to the ARR Code of Ethics, particularly with respect to transparency, reproducibility, and responsible computing practices.

Acknowledgements

The research was supported by Shanghai Artificial Intelligence Laboratory, a locally commissioned task from the Shanghai Municipal Government, the Shanghai Municipal Science and Technology Major Project, and the Youth Talent Research Funding Program (Grant No. P25RZ00020HJ).

References

2025. Manus. https://manus.im/.

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, and 1 others. 2025. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*.

OpenAI / SWE bench Team. 2024. Swe-bench verified: A human-validated subset for robust evaluation of ai coding agents. https://openai.com/index/introducing-swe-bench-verified/.

Yuxuan Cai, Yipeng Hao, Jie Zhou, Hang Yan, Zhikai Lei, Rui Zhen, Zhenhua Han, Yutao Yang, Junsong Li, Qianjun Pan, and 1 others. 2025. Building self-evolving agents via experience-driven lifelong learning: A framework and benchmark. *arXiv preprint arXiv:2508.19005*.

Jingyi Chai, Shuo Tang, Rui Ye, Yuwen Du, Xinyu Zhu, Mengcheng Zhou, Yanfeng Wang, Yuzhi Zhang, Linfeng Zhang, Siheng Chen, and 1 others. 2025. Scimaster: Towards general-purpose scientific ai agents, part i. x-master as foundation: Can we lead on humanity's last exam? *arXiv preprint arXiv:2507.05241*.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision (ECCV)*.

Google. 2024. Introducing gemini deep research. https://blog.google/products/gemini/google-gemini-deep-research/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Bernard Ghanem, Ping Luo, and Guohao Li. 2025. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. *Preprint*, arXiv:2505.23885.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, and 1 others. Deep research agents: A systematic examination and roadmap, 2025b. *URL https://arxiv. org/abs/2506*, 18096.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

Jiarun Liu, Shiyue Xu, Shangkun Liu, Yang Li, Wen Liu, Min Liu, Xiaoqing Zhou, Hanmin Wang, Shilin Jia, Shaohua Tian, and 1 others. 2025. Joyagent-jdgenie: Technical report on the gaia. *arXiv preprint arXiv:2510.00510*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. *arXiv preprint arXiv:2401.13178*.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations (ICLR)*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2025a. Chatgpt (gpt-5).

OpenAI. 2025b. Deep research system card. https://cdn.openai.com/deep-research-system-card.pdf.

Perplexity. 2025. Perplexity deep research. https://www.perplexity.ai/.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, and 1 others. 2025. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *arXiv preprint arXiv:2509.13309*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems (NeurIPS)*.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Zhili Shen, Chenxin Diao, Pavlos Vougiouklis, Pascual Merita, Shriram Piramanayagam, Enting Chen, Damien Graux, Andre Melo, Ruofei Lai, Zeren Jiang, Zhongyang Li, Qi Ye, Yang Ren, Dandan Tu, and Jeff Z. Pan. 2024. Gear: Graph-enhanced agent for retrieval-augmented generation. *arXiv preprint arXiv:2412.18431*.

Wenxuan Shi, Haochen Tan, Chuqiao Kuang, Xiaoguang Li, Xiaozhe Ren, Chen Zhang, Hanting Chen, Yasheng Wang, Lifeng Shang, Fisher Yu, and Yunhe Wang. 2025. Pangu deepdiver: Adaptive search intensity scaling via open-web reinforcement learning. *arXiv preprint arXiv:2505.24332*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, and 1 others. 2025. Webshaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*.

MiroMind AI Team. 2025. Miroflow: A high-performance open-source research agent framework. https://github.com/MiroMindAI/MiroFlow.

NovelSeek Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, and 1 others. 2025a. Novelseek: When agent becomes the scientist–building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025b. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, and 5 others. 2025. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations (ICLR)*.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, and 1 others. 2025. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.

xAI. 2025. Grok-3 deepsearch: Synthesizing key information to distill clarity from complexity. https://x.ai/news/grok-3.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yue Dong. 2023b. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2303.11366*.

Chengyue Yu, Siyuan Lu, Chenyi Zhuang, Dong Wang, Qintong Wu, Zongyue Li, Runsheng Gan, Chunfeng Wang, Siqi Hou, Gaochi Huang, Wenlong Yan, Lifeng Hong, Aohui Xue, Yanfeng Wang, Jinjie Gu, David Tsai, and Tao Lin. 2025. Aworld: Orchestrating the training recipe for agentic ai. *arXiv preprint arXiv:2508.20404*.

Zhongyue Zhang, Zijie Qiu, Yingcheng Wu, Shuya Li, Dingyan Wang, Zhuomin Zhou, Duo An, Yuhan Chen, Yu Li, Yongbo Wang, and 1 others. 2025. Origene: A self-evolving virtual disease biologist automating therapeutic target discovery. *bioRxiv*.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. 2024. Gptswarm: Language agents as optimizable graphs. In *International Conference on Machine Learning (ICML)*.

A Tools of Knowledge Collector

We provide a set of tool wrappers used by the Knowledge Collector. Each tool is designed with concurrent safety, creating independent toolkit instances to avoid state conflicts. Table 3 summarizes the available tools.

Table 3: The tools in Knowledge Collector

Tool	Purpose
`search_google`	Use Google search engine to search information for the given query
`search_wiki`	Search the entity in WikiPedia and return the summary of the required page, containing factual information about the given entity
`search_wiki_revision`	Search Wikipedia to get the latest Wikipedia revision *at or before* the end of the given (year, month)
`search_archived_webpage`	Given a url, search the wayback machine and returns the archived version of the url for a given date
`extract_document_content`	Extract the content of a given local document and return the processed text. It can process various types of documents, including text, image, table, audio, video, zip, json, xml, pdf, py etc
`extract_url_content`	Extract the html content of a given url and return the processed text
`ask_question_about_image`	Answer image questions with optional custom instructions
`ask_question_about_audio`	Ask a question about the audio and get the answer using multimodal model
`ask_question_about_video`	Ask a question about the video using Gemini multimodal capabilities
`download_media_from_url`	Download any given URL (image, video, audio, document, or webpage)
`execute_code`	Execute a given code snippet
`browse_url`	A powerful toolkit which can simulate the browser interaction to solve the task which needs multi-step actions
`ocr2text`	OCR the image and return the text

B Conclusion Generation

A summarizer for conclusion generation has been developed to generate conclusions for the answer node, featuring two modes of operation.

Question-Answering Tasks: When the objective is to answer a specific question, the task usually requires a strict logical progression of reasoning. In such cases, the later nodes in the execution graph—particularly solve nodes—tend to encapsulate the reasoning steps that are most directly related to the final answer. By contrast, earlier nodes such as search nodes often contain intermediate evidence or auxiliary information that, while necessary for the reasoning process, does not itself contribute to the correctness or clarity of the final response. To enhance efficiency and reduce noise, the summarizer in this mode selectively incorporates only the dependent predecessor nodes of the final answer. This targeted approach ensures that the summary remains focused, concise, and aligned with the logical chain that directly supports the solution. The main benefit is an improvement in answer precision and interpretability, as irrelevant or redundant details are filtered out.

Report-Generation Tasks: By contrast, when the task involves producing a comprehensive report, the goal is not merely accuracy but also coverage and richness. In this context, limiting the summarizer to dependent nodes would risk omitting potentially valuable background, context, or supporting evidence. Therefore, for report generation, the entire execution graph produced by FlowSearch is passed to the summarizer. This design allows the system to synthesize information from all nodes—including search, solve, and answer stages—so that the final report captures not only the core reasoning steps but also the broader landscape of evidence. The benefit of this approach is that the generated report provides a more holistic view of the research process, offering readers both the conclusions and the supporting context, which increases transparency, interpretability, and informational richness.

Advantages of the Dual-mode design: This bifurcated summarization strategy balances efficiency with completeness. For question answering, it minimizes cognitive load and reduces error propagation by concentrating only on essential reasoning chains. For report writing, it maximizes informativeness and ensures that potentially useful evidence is not prematurely discarded. Together, these modes enable FlowSearch to flexibly support both precise problem-solving and broad knowledge synthesis, depending on the user's research goals.

C Planner ablation

To make an attempt at the training of the base model, we employ knowledge distillation from GPT-o4-mini to train a specific planner for FlowSearch. Specifically, for the training data, we collect a set of Wikipedia entries that inherently exhibit entity dependencies. These dependencies are extracted and organized into structured entity graphs. The entity graphs are then fed into o4-mini, which generates questions for each node based on its corresponding dependencies and subsequently integrates these questions into a single summary question, serving as the user query in our dataset. For the graph generation process, the obtained entity graphs are directly converted into a natural language representation according to our predefined format, which is included as the assistant output in the labeled dataset. A detailed description of the data format is provided in Box C, and the results are shown in Table 4.

During training, the labeled data is decomposed into multiple single-turn question-answer pairs, which are then used to fine-tune Qwen-3-series models via supervised fine-tuning (SFT).

Table 4: Ablation study on the planner model. We compare various flow planners, including the Qwen3 series and our finetuned Trained-Planner. Results are reported on the GAIA benchmark.

	GAIA			
Planner	Level 1	Level 2	Level 3	Avg
Qwen-3-8B	58.49	46.51	11.54	44.85
Trained-Planner-8B (ours)	70.25	67.44	34.61	66.06
Qwen-3-32B	77.36	67.44	30.77	64.81
Trained-Planner-32B (ours)	**84.91**	**70.93**	**42.31**	**70.91**

C. Data Format

```
{
  "messages": [
    {
      "role": "user",
      "content": "You are a graph
          planner agent. Your task
          is to decompose any user
          question into a logical
          graph of tasks, and
          iteratively refine the
          graph when node knowledge
           becomes available.

        ### Example Input Graph
        {
          "nodes": [
            {"node_id": "n1", "
                type": "answer", "
                task": "Explain why
                 sugar-free drinks
                can still contain
                carbohydrates"}
          ],
          "edges": []
        }

        Input graph:
        <generated_input_graph>"
    },
    {
      "role": "assistant",
      "content": "<
          generated_output_graph>"
    }
  ]
}
```

D Efficiency Analysis

We also report a resource comparison in Table 5. FlowSearch achieves favorable cost and latency trade-offs primarily due to its flow-driven design, which enables efficient orchestration, parallel execution, and early termination when appropriate. While a cost-effective base model is used, the observed efficiency gains mainly stem from the system-level design rather than model choice alone.

Table 5: The resource comparison between OWL and FlowSearch on the GAIA benchmark.

Method	API Cost ($/query)	Time (s/task)
OWL	1.20	236
FlowSearch	0.86	177

E Additional Case Studies

E.1 Question Answering

Below we choose one question from GAIA level 2 to show our execution logic. The content of the question is **"What is the latest chronological year date in the image from the first citation of Carl Nebel's Wikipedia page (Aug 2023)?"**. This question is inherently *logic-intensive*: the answer is not present on the query page but must be derived through a chained, evidence-preserving procedure.

Our FlowSearch solves this question step by step, in a very logical order. Starting from the revision-resolved entry point, the planner instantiated a dependency-aware, tool-grounded pipeline: `n1` resolves the August 2023 Wikipedia revision; `n2` extracts the first citation URL; `n3` fetches the citation page HTML; `n8` identifies and downloads the first in-article image; `n6` performs OCR over the downloaded image; and `n7` parses four-digit years to determine the latest date. This design operationalizes a search–extract–process–analyze flow where each step produces verifiable intermediate artifacts (URLs, HTML snapshots, files, OCR text) that can be audited, cached, and reused. By encoding dependencies in a graph, FlowSearch enables deterministic, provenance-preserving re-execution, isolates errors to specific nodes, and supports targeted recovery without rerunning the entire pipeline—yielding stronger reproducibility, interpretability, and multi-tool coordination than monolithic, end-to-end prompting. Table 6 shows every node in our FlowSearch graph system, includ-

ing task content, type, status and its output(part). Since this question requires each step to rely strictly on the results of the previous one, the overall process forms a linear execution flow. The execution followed a strict topological ordering:

$$\text{n1} \rightarrow \text{n2} \rightarrow \text{n3} \rightarrow \text{n8} \rightarrow \text{n6} \rightarrow \text{n7} \rightarrow \text{task}$$

E.2 Report Generation

The following report was generated by our FlowSearch system to answer the query **"Help me research the latest progress in multi-agent AI scientists in 2025"**. The planner decomposed the original research query into a set of interconnected subtasks, enabling systematic information gathering, reasoning, and synthesis. The resulting execution graph consisted of 7 nodes spanning three major categories—search, solve, and answer—which together captured the full problem-solving workflow. Table 7 provides a structured overview of all nodes and their corresponding roles within the pipeline. Importantly, in the case of report generation, the graph structure highlights its advantages even more clearly, since the dependencies across different report sections are relatively weak, allowing parallel execution to be fully leveraged. Therefore, we can generate the report below in 10 minutes. An example output report is shown below.

The execution followed a topological ordering, where nodes grouped together indicate parallel execution:

$$\text{n3, n4s, n7} \rightarrow \text{n2, n4, n6} \rightarrow \text{task}$$

Table 6: Execution Trace of FlowSearch(Question Answering Case Study)

Node	Task	Type	Status	Output(Part)
n1	Use `search_wiki_revision` to get Carl Nebel Wikipedia revision (Aug 2023)	search	Success	The Carl Nebel Wikipedia page revision dated 2023-08-05T13:53:28Z carries oldid 1168855983 and is accessible at https://en.wikipedia.org/w/index.php?title=Carl_Nebel&oldid=1168855983
n2	Open Carl Nebel revision and extract first citation URL from References	solve	Success	The first citation in the References section is 'Thieme–Becker', entry "Nebel, Carl," with URL https://de.wikipedia.org/wiki/Thieme-Becker
n3	Use `extract_url_content` to fetch citation page HTML	search	Success	Extracted the HTML content of the German Wikipedia page at https://de.wikipedia.org/wiki/Thieme-Becker, which presents Thieme-Becker as a German biographical dictionary of artists...
n8	Extract first image from citation page and download locally	search	Success	The first image in the 'Thieme-Becker' article body has source URL https://upload.wikimedia.org/wikipedia/commons/thumb/c/c5/Perwanger%2C_Chr...
n6	Apply `ocr2text` on downloaded image	solve	Success	The OCR2Text tool returned success and extracted the following lines of text: Pervinquiere-Pescatori; Pervinquiere, Henri Baron, Tiermaler,; Perz, Michael, Stukkator, tatig 1701 im; 1883 Poitiers; Bild im Mus. ebda.; Rathaus zu Landsberg a. Lech...
n7	Parse OCR text to extract all 4-digit years and find latest	solve	Success	From the OCR-extracted text, the unique four-digit years identified are [1558, 1577, 1610, 1645, ... 1913, 1915, 1924, 1925, 1927], and the latest chronological year among these is 1927.
task	What is the latest chronological year date in the image from the first citation of Carl Nebel's Wikipedia page (Aug 2023)?	answer	Success	1927

Table 7: Execution Trace of FlowSearch(Report Generation Case Study)

Node ID	**Type**	**Task Description**
task	answer	Help me research the latest progress in multi-agent AI scientists in 2025.
n2	solve	[Background & Methods] Synthesize definitions, scope, historical context, and classify core methods
n3	search	[Background & Methods] Collect definitions, seminal works, and representative methods
n4	solve	[Datasets/Applications] Summarize datasets, benchmarks, evaluation metrics, and applications
n4s	search	[Datasets/Applications] Collect datasets, benchmarks, evaluation results, and application examples
n6	solve	[Challenges & Future Work] Analyze challenges, limitations, and outline future directions
n7	search	[Challenges & Future Work] Collect discussions of current challenges, limitations, and future outlook

F Prompts

In this section, we present the prompts used by each agent to clarify the workflow of FlowSearch. The prompt design is closely aligned with the proposed knowledge flow-based planning paradigm, rather than a sequential planner. Specifically, the prompts explicitly encode node-level responsibilities, dependency-aware execution, and intermediate state passing, which are essential for supporting non-linear research trajectories in the planner.

In addition, the refiner prompt reflects a prior that intermediate results produced by the graph planner may be incomplete or inconsistent, and thus require global consolidation and revision. These prompts therefore not only specify agent behaviors, but also serve as an explicit interface between the graph planner and the refinement stage in FlowSearch.

Flow Planner Prompt

You are a graph planner agent.
Decompose any user question into a logical graph of tasks, refining iteratively when node knowledge becomes available.

Output strictly JSON with "nodes" and "edges".

Node rules:
- nodes:
  - "node_id": unique id (e.g., "n1")
  - "type": ["search", "solve", "answer"]
  - "task": short natural language description
- edges:
  - "from", "to", "relationship"

Node type:
- search: collect raw info
- solve: reason, compute, integrate, or handle complex tasks
- answer: final solution (only one)

Refinement:
- Break nodes into concrete child tasks and connect edges
- Add edges if a node depends on another's knowledge
- Modify incorrect/unreasonable tasks
- Expand only one layer per iteration
- Stop and output "Perfect!" if all nodes are concrete and complete, and no further decomposition is possible
- For survey/review questions, ensure major aspects/perspectives are covered

Input format:
- JSON graph with initial "answer" node
- Do not modify answer node
- Always produce valid JSON
- If nodes > max nodes, do not add new nodes, clarify existing ones; if clear, output "Perfect!"

Example behavior:
- If all nodes concrete and sufficient → "Perfect!"
- Otherwise, add one layer of concrete child nodes

Example

Input graph:
```json
{{
  "nodes": [
    {{"node_id": "n1", "type": "answer", "task": "Explain why sugar-free drinks can still contain carbohydrates"}}
  ],
  "edges": []
}}
```
Make sure to finish your plan in {max_iter} turns, and this is your {current_iter} turn.

Make sure to not add more than {max_nodes} nodes.

This is the input graph {graph} to answer the question{question}

Flow Refiner Prompt

You are a Graph Reasoning Agent managing and updating DAG task graphs for multi-step workflows.

You are given a graph and a query, and need to modify the graph to answer the query.

Input
1. **graph**: JSON with
   - `nodes`: each with `node_id`, `status` (pending/executed), `task`, `type` (search/solve/answer), `final_response`, `success`, `reasoning`
   - `edges`: each with `from`, `to`, `relationship`
2. **query**: overall problem the graph solves

Only pending nodes can be modified. Do not modify executed nodes or the answer node.

Allowed Actions
- Node: `add_node`, `remove_node`, `modify_node`
- Edge: `add_edge`, `remove_edge`, `modify_edge`

Modification Rules
- Refine unclear tasks, remove redundant nodes, add nodes for alternative execution paths.
- Add/remove edges to maintain correct dependencies.
- Keep the graph connected, acyclic, and minimal.
- Only modify nodes/edges needed to fix failures.
- If no changes are needed, output `[]`.
- Do not change the answer node.

Output Format
JSON array of modifications. Each modification includes:
- `action`: one of allowed actions
- `node_id` / `from_node` / `to_node`
- `attributes`:
  - Node: `{ "task": "...", "type": "search|solve|answer"}`
  - Edge: `{ "relationship": "..." }`
- `reason`: concise explanation

Example Output
```json
[
 {
  "action": "add_node",
  "node_id": "n6",
  "attributes": {
   "task": "Validate the final answer against multiple sources",
   "type": "solve"
  },
  "reason": "Introduce an explicit validation step to improve reliability"
 },
 {
  "action": "add_edge",
  "from_node": "n3",
  "to_node": "n6",
  "attributes": { "relationship": "produces draft answer" },
  "reason": "The output of n3 should flow into the new validation step"
 },
 {
  "action": "add_edge",
  "from_node": "n6",
  "to_node": "n4",
  "attributes": { "relationship": "validated answer" },
  "reason": "Ensure the validated result is passed downstream to n4"
 }
]
```
This is the input graph:

==========
{graph}
==========

This is the query that the graph is meant to solve:

==========
{query}
==========

Knowledge Collector Prompt

You are a Task Graph Agent. Your job is to:
1) Decompose the given task into minimal, concrete subtasks.
2) Execute subtasks using available tools or upstream knowledge.

Inputs
- **task**: current task description
- **query**: overall question
- **Full subtask list**: history subtask list
- ** subtask **: Current subtask to execute
- **tools**: available execution tools

Subtask Planning Rules
- Split the task into as few subtasks as possible (<{max_subtasks}>), unless it is a survey/review task, then cover major aspects.
- Subtasks describe **what to do**, not how. No tools, parameters, or implementation details in subtask text.
- Return subtasks as:
<tasks>
<task>Subtask 1</task>
<task>Subtask 2</task>
</tasks>

Execution Rules
- Decide for each subtask whether to use tools or answer directly.
- Always use appropriate tools for knowledge retrieval, code execution, or document processing.
- Include concise, reproducible outputs; list files produced.
- Reference upstream `knowledge_info` to avoid redundant work.
- If blocked by missing inputs or errors, return a clear blocker description.
- Do NOT reveal chain-of-thought. Keep outputs machine-readable and self-contained.

Output
- For planning: numbered list of subtasks as above
- For execution: structured JSON with results, artifacts, and any notes on blockers
"""
Overall task:
==========
{task}
==========

The task serves as a part to solve this question(- This only tells you what the overall question is. - You are NOT asked to solve the query directly.):
==========
{query}
==========
Full subtask list (from the planner), ordered, some subtasks are already completed:
==========
{history_subtasks}
==========

Current subtask to execute (one item from the list above and this is your current subtask):
==========
{subtask}
==========
Here are the available tools

==========
{tools}
==========

The following figures show an example of a report generated by FlowSearch:

The State of Multi-Agent AI Scientists: A 2025 Progress Report on Applications, Challenges, and Future Trajectories

Introduction

The application of artificial intelligence to scientific discovery is undergoing a significant paradigm shift, moving from passive computational tools toward integrated systems of collaborating autonomous agents known as "multi-agent AI scientists" [[1]][[2]][[3]]. As of 2025, this field has transitioned from primarily theoretical concepts to practical applications, with agentic frameworks now being deployed to accelerate research in high-stakes domains such as biomedical science, quantum materials, and climate science [[4]][[5]][[6]]. Systems like Google's AI Co-Scientist and the MASTER framework for materials discovery exemplify this trend, demonstrating the capacity to autonomously manage complex workflows from literature review and hypothesis generation to experimental design and data analysis [[4]][[7]]. These platforms function as proactive collaborators, augmenting human research teams by tackling problems of immense scale and complexity [[8]][[9]].

Despite this rapid progress, the development of robust and truly autonomous AI scientists is confronted by significant challenges. Key technical hurdles include achieving strategic coordination and collaborative reasoning among agents, orchestrating complex multi-step workflows, and ensuring that individual agents possess grounded, domain-specific scientific reasoning capabilities [[10]][[11]][[12]]. Alongside these technical obstacles are pressing ethical considerations. The increasing autonomy of these systems raises critical questions regarding accountability for erroneous results, the potential for dual-use applications, the integrity of AI-generated knowledge, and the socio-economic and environmental impacts of their deployment [[1]][[3]][[6]][[13]].

This report provides a comprehensive overview of the state of multi-agent AI scientists in 2025. It begins by establishing a foundational definition of the paradigm, its core capabilities, and its primary architectural models. The report then examines current real-world applications and breakthroughs, followed by an analysis of the innovation ecosystem of key institutions and projects. Subsequently, it offers a detailed discussion of the field's primary limitations and ethical challenges. The report concludes by projecting the future trajectory of research, including emerging technical capabilities and expanding application frontiers.

Defining the Paradigm: Capabilities and Architectures of AI Scientists

The systems known as 'multi-agent AI scientists' represent a specialized form of multi-agent systems (MAS) purpose-built to conduct end-to-end scientific research with minimal human intervention. A multi-agent AI scientist can be defined as a computational system composed of multiple, distinct, and interacting intelligent agents that collaboratively operate within a shared environment to execute complex scientific workflows. The foundational principle is the distribution of tasks and communication among these agents, enabling a structured, parallel, and specialized approach to problem-solving that emulates the collaborative dynamics of a human research team [[1]][[2]][[3]]. The constituent agents are designed to be autonomous, capable of independent decision-making, learning from experience, and adapting their behavior to achieve both individual sub-goals and a collective research objective.

This paradigm is fundamentally distinct from both single-agent AI researchers and traditional scientific software. Whereas a single-agent system operates as a monolithic cognitive entity with a centralized decision-making process, a multi-agent system introduces the complex dynamics of interaction, coordination, and cooperation among multiple autonomous actors. The focus shifts to the inter-agent social and strategic dimensions that are absent in its single-agent counterpart [[10]][[11]]. Likewise, these systems signify a qualitative leap beyond traditional scientific software, such as simulators or statistical packages, which function as passive instruments requiring explicit, step-by-step human instruction. The defining characteristic of the AI scientist is its autonomy—the ability to proactively navigate the entire research lifecycle, from initial ideation to the preparation of publication-ready manuscripts. This transforms AI from a mere tool into a proactive collaborator in the discovery process [[14]].

A Hierarchy of Core Capabilities

The efficacy of a multi-agent AI scientist system is contingent on an integrated suite of capabilities spanning foundational cognitive functions, scientific workflow competencies, and advanced collaborative mechanisms.

At the most fundamental level, individual agents must possess robust cognitive and information processing abilities. This includes perceiving and interpreting diverse data formats, such as text and images, to extract salient features for reasoning [[15]]. Generative AI provides agents with the crucial dual capabilities of **information generation**—creating novel content like research summaries—and **information synthesis**. The latter, often operationalized through techniques like Retrieval-Augmented Generation (RAG), allows agents to integrate and reorganize knowledge from vast scientific literature. This function is critical for producing grounded, factually accurate outputs and mitigating the risk of model hallucination [[16]].

Building upon this cognitive foundation are the competencies that map directly onto the scientific method. These include:

- **Hypothesis Generation:** Agents analyze existing knowledge to formulate novel, testable research questions and construct coherent proposals [[4]][[17]].
- **Experimental Design and Planning:** This involves translating an abstract hypothesis into a concrete, executable protocol. This capability can range from designing virtual simulations to planning complex real-world procedures, as demonstrated by AI systems that automate intricate radiotherapy treatment plans by handling steps like beam setup, optimization, and plan quality improvement [[18]][[17]].
- **Execution and Analysis:** The workflow continues with the capacity for coordinating specialized software for experimental execution, followed by **data analysis**, where agents interpret the outcomes to draw conclusions and refine their understanding [[19]].

Perhaps the most defining characteristic of the multi-agent paradigm is the set of advanced capabilities that enable effective coordination and collaboration. For a collective of specialized agents to function as a cohesive team, they must be able to **communicate, collaborate, and learn together** to achieve complex, overarching goals [[20]]. A key challenge in frontier research is managing cooperation under conditions of partial observability and limited communication. One sophisticated solution involves establishing shared **"conventions"**—predefined principles or rules that augment the agents' standard action space. These conventions enable them to convey complex ideas and coordinate actions over multiple time steps, thereby improving cooperative performance in a manner analogous to human expert teams [[11]]. Furthermore, the collective's effectiveness often depends on **sequential and strategic decision-making**, where the precise order of agent actions is paramount. Advanced Multi-Agent Reinforcement Learning (MARL) models, such as the Agent Order of Action Decisions-MAT (AOAD-MAT), explicitly address this by learning to predict and dynamically adjust the optimal sequence of agent actions to maximize the team's synergistic advantage [[10]][[21]].

Key Architectural Paradigms for Realization

The implementation of multi-agent AI scientists is critically dependent on the underlying architecture, which dictates the patterns of communication, coordination, and control. The choice of paradigm influences the system's scalability, robustness, and adaptability, with models ranging from classical centralized structures to dynamic, learning-based orchestrators [[22]].

One fundamental distinction is between **centralized and decentralized coordination**. A centralized architecture employs a master agent or orchestrator for global planning and task allocation. While this simplifies coordination, the central controller can become a computational bottleneck and represents a single point of failure, limiting flexibility and resilience [[23]]. In contrast, a decentralized architecture distributes control and decision-making across all agents, which coordinate via peer-to-peer communication. This model is inherently more robust and scalable but makes achieving coherent collective behavior more complex, as each agent operates with only partial observability. A significant challenge in this paradigm is the "straggler effect" caused by heterogeneity in agent resources, particularly in computationally intensive science. Frameworks like Communication-Efficient Training Workload Balancing for Decentralized Multi-Agent Learning (ComDML) address this by enabling slower agents to offload workloads to faster ones, balancing computation and communication to enhance overall efficiency [[24]].

A more sophisticated hybrid model is the **blackboard system**, which is particularly well-suited for complex, ill-defined problems like scientific discovery. This architecture consists of three core components: a set of independent specialist agents known as **Knowledge Sources (KSs)**; a global, shared data repository called the **blackboard** where problems and partial solutions are posted; and a **control shell** that moderates agent access to the blackboard [[25]]. This structure facilitates an opportunistic and incremental problem-solving process. For instance, a 'Hypothesis Generator' agent posts a theory to the blackboard, which might trigger an 'Experiment Designer' agent to post a protocol, which is then executed by a 'Simulation' agent that posts results for a

'Data Analyst' agent to interpret. This fluid workflow, where progress emerges from collective contributions, replaces rigid task assignment and has been explored for use in modern LLM-based systems to enable more dynamic communication [[23]][[25]].

At the forefront of architectural design is **reinforcement learning-based orchestration**. In this highly dynamic paradigm, the coordination strategy is not hard-coded but is learned through MARL as agents seek to maximize a collective reward. A critical insight driving this research is that the order of agent actions is often paramount to success [[10]]. Advanced Transformer-based models like the aforementioned AOAD-MAT use an actor-critic architecture to explicitly learn and predict the optimal order of agent actions for a given task, effectively creating a learned, adaptive orchestration policy. This approach moves beyond static architectures by enabling the system to discover and refine its own coordination protocols. Further research in this area explores dynamic team formation, developing frameworks where agents can learn to form bilateral teams algorithmically within dynamic populations, adding another layer of architectural flexibility [[10]][[26]]. Ultimately, the optimal architecture for a given multi-agent AI scientist is context-dependent, contingent on the scientific domain and the desired balance between explicit control and emergent, autonomous discovery.

Current Applications and Breakthroughs in 2025

The year 2025 has emerged as a significant inflection point for the application of artificial intelligence in scientific research, characterized by the growing adoption of multi-agent AI systems as collaborative partners. This paradigm, often referred to as "Agentic Research," marks a departure from single-purpose AI tools toward complex ecosystems of autonomous, specialized agents that collectively manage the entire scientific process [[8]][[9]]. The development of these systems has been significantly accelerated and democratized by the availability of enabling platforms, such as OpenAI's "ChatGPT Agent" toolbox [[13]].

Expanding Frontiers: New Domains of Application

Multi-agent AI systems are proving their value across several high-stakes scientific fields, where their unique architectural capabilities are well-suited to address long-standing challenges.

Quantum Materials and Computational Chemistry
The discovery of novel materials, particularly quantum materials, is a field constrained by a vast, high-dimensional search space, the slow pace of experimental synthesis, and the intensive computational demands of simulations. Multi-agent AI frameworks directly confront these bottlenecks through an orchestrated architecture that partitions the research workflow [[5]]. For instance, a system can be composed of a literature-parsing agent to identify promising material classes, a simulation agent to execute quantum chemistry calculations, an interpretation agent to predict material properties from simulated data, and a synthesis agent to propose viable production pathways [[5]]. This division of labor not only accelerates the research cycle but also democratizes access to advanced computational methods. As demonstrated in several 2025 preprints, these agentic tools encapsulate the complexity of quantum chemistry calculations, thereby lowering the barrier to entry for scientists who lack deep expertise in computational physics [[27]]. The long-term vision in this domain is the creation of fully integrated, AI-assisted laboratories where agentic systems design materials and interface with robotic platforms to autonomously synthesize and test them, establishing a closed-loop discovery engine that is dramatically faster and more cost-effective [[28]].

Drug Discovery and Biomedical Science
The pharmaceutical sector, historically defined by long and costly research and development cycles, has become a primary beneficiary of multi-agent AI. The core challenges in this field include navigating the immense body of biomedical literature to identify novel biological targets, generating testable hypotheses about disease mechanisms, and designing effective, low-toxicity drug candidates [[4]]. In 2025, multi-agent platforms powered by advanced foundation models like Gemini 2.0 are functioning as virtual "AI co-scientists." These systems can autonomously generate innovative hypotheses and structure comprehensive research proposals that might otherwise be overlooked, allowing human research teams to explore a wider intellectual landscape and better prioritize their efforts [[4]]. A significant trend is the development of open and collaborative AI drug discovery pipelines, which invite broad engagement from clinical, translational, and computational experts to refine and validate the models. As highlighted at major AI science conferences, this represents a shift away from proprietary, black-box systems toward a more verifiable, community-driven process [[29]]. The tangible impact is already evident, with AI-driven platforms contributing to a portfolio of active development programs for oncology and rare diseases that are populating the real-world clinical pipeline by mid-2025 [[30]].

Climate Science and Earth System Modeling
Climate science is fundamentally dependent on computationally intensive numerical models, ranging from simpler Energy

Balance Models (EBMs) to complex General Circulation Models (GCMs) and Earth System Models (ESMs), which simulate the intricate interactions between the planet's atmosphere, oceans, land, and cryosphere using systems of differential equations derived from physical laws [[6]]. Key challenges in this field include the immense computational cost and electricity consumption of running high-resolution simulations on exascale supercomputers, as well as the uncertainties introduced by parametrization (the representation of small-scale processes like cloud formation). While direct applications are still emerging, the structural challenges of climate science align perfectly with the capabilities of agentic systems. A multi-agent framework could be designed to manage ESM complexity, with individual agents overseeing sub-models (e.g., an 'Ocean Agent' or 'Atmosphere Agent') and coordinating their data exchanges, mirroring the orchestration seen in materials science [[5]][[6]]. Such a system could automate computationally demanding tasks like "extreme event attribution" by autonomously designing and executing thousands of model runs to quantify the impact of anthropogenic warming on specific weather events. Furthermore, AI agents could be tasked with optimizing simulation performance by developing novel algorithms to reduce floating point precision or by using machine learning to avoid redundant calculations, thereby addressing the critical issue of electricity consumption and making high-fidelity climate science more sustainable [[6]].

Case Studies: From Theory to Tangible Success in 2025

The practical impact of agentic research is best illustrated by several high-profile deployments that have produced tangible scientific breakthroughs.

Case Study: Google's AI Co-Scientist in Biomedical Research
One of the most prominent demonstrations is Google's AI Co-Scientist, a multi-agent platform designed to accelerate the initial, ideation-heavy phases of biomedical research [[4]]. The system's primary objective is to autonomously survey scientific literature, generate novel and testable hypotheses, and structure comprehensive research proposals for human scientists to pursue [[4]]. Built upon the Gemini 2.0 foundation model [[31]], the platform operates as a collaborative ecosystem of specialized agents, likely including a Literature Analysis Agent, a Hypothesis Generation Agent, and a Proposal Structuring Agent. This architecture embodies the "AI co-scientist" paradigm, where the AI performs the cognitive labor of ideation and literature grounding alongside human experts [[4]]. In a specific 2025 application, the platform was tasked with identifying novel drug candidates with significant anti-fibrotic properties [[32]]. The results were transformative: the system replicated a decade of human-led experimental insight in just two days and reduced the hypothesis generation phase from months to mere days [[32]][[33]]. Crucially, the AI identified promising drug candidates that subsequently showed significant anti-fibrotic activity in validation experiments [[32]]. This success validates that multi-agent systems can overcome human cognitive bottlenecks by reasoning over a vast knowledge corpus to uncover non-obvious connections, confirming that the co-scientist model—where AI generates hypotheses and humans drive validation—is a powerful and viable path forward for biomedical discovery [[4]][[33]].

Case Study: The MASTER Framework for Autonomous Materials Discovery
The MASTER (Materials Agents for Simulation and Theory in Electronic-structure Reasoning) framework represents a significant advance toward fully autonomous scientific reasoning in materials science [[7]]. Its objective is to achieve true autonomy in discovering functional materials by independently designing, executing, and interpreting complex Density Functional Theory (DFT) simulations without human intervention in the reasoning loop [[7]]. This work advances the concept of "inverse design," where materials are generated to meet specific desired properties [[34]]. MASTER's hierarchical, multi-agent architecture features a lower-level **Multimodal Translation Agent** that converts high-level objectives into executable simulation workflows, and a higher-level team of **Reasoning Agents** that guide the discovery process using collaborative strategies. These strategies include **Peer Review**, where agents critique each other's proposed experiments, and **Triage-Ranking**, used to prioritize the most promising simulations [[7]]. In challenging chemical discovery problems, such as optimizing CO adsorption on catalysts, MASTER's reasoning-driven exploration reduced the number of computationally expensive atomistic simulations required by up to 90% compared to baseline methods [[7]]. This massive efficiency gain addresses a primary bottleneck in computational materials science. The project provides compelling evidence that a multi-agent LLM framework can engage in chemically grounded scientific reasoning far more sophisticated than simple pattern matching. The hierarchical structure and the analysis of collaborative strategies offer critical insights into designing more effective AI scientist teams, marking a paradigm shift toward truly autonomous scientific exploration [[7]][[34]].

Case Study: AI-Driven Design of Protein Binders for Cancer Immunotherapy
A 2025 collaboration between the Technical University of Denmark and Scripps Research resulted in an AI platform that can rapidly design custom proteins for cancer immunotherapy [[13]]. The platform's goal was to drastically shorten the development timeline for novel protein minibinders, which are engineered to help a patient's own T-cells identify and destroy cancer cells. While described as an integrated "AI platform," its complexity implies a modular, agentic workflow involving a **Target Identification Agent**, a **Generative Design Agent**, a **Binding Simulation Agent**, and a **Therapeutic Efficacy Agent**, mirroring

the division of labor seen in other scientific domains [[13]]. The platform achieved an unprecedented speed, moving from concept to a validated design in weeks—a process that traditionally takes years. The AI-generated protein designs were successfully synthesized and, in subsequent lab experiments, were shown to guide T-cells to selectively target and kill cancer cells. This represents a direct and rapid translation from in-silico design to functional therapeutic action [[13]]. This case study is a powerful illustration of a closed-loop discovery cycle, where AI-driven design is rapidly validated by experiment. The key insight is the platform's ability to navigate the astronomically large possibility space of protein sequences to find functional, stable molecules with remarkable speed, effectively bridging the gap between computational biology and practical medicine [[13]].

The Innovation Ecosystem: Key Institutions, Labs, and Projects

The landscape of artificial intelligence in 2025 is undergoing a fundamental redefinition, marked by a strategic shift from monolithic, single-purpose models to dynamic, collaborative multi-agent systems. This era, dubbed the "Year of the AI Agent," represents a significant conceptual and technical evolution, unlocking new potentials for autonomous and collective problem-solving [[35]]. The progress in this domain is not the result of a single entity's efforts but rather the output of a vibrant, symbiotic ecosystem. This ecosystem is composed of three interdependent pillars: premier academic and research institutions that provide foundational theory, corporate laboratories that pioneer large-scale applications and commercialization, and a robust open-source community that builds the infrastructure to democratize these powerful new capabilities.

Academic and Research Foundations

The theoretical and conceptual underpinnings of multi-agent AI are being forged within a distributed, global network of research institutions. In North America, Stanford University's Institute for Human-Centered AI (HAI) has been a key force, with its analyses forecasting the strategic migration from individual AI models to complex, collaborative agent systems designed to emulate human expert teams [[36]]. This thought leadership is supported by comprehensive empirical resources like the annual AI Index Report, which provides data-driven synthesis for both academic and policy-making audiences [[37]]. Carnegie Mellon University (CMU) makes significant contributions through its longstanding strengths in core AI disciplines, including robotics, planning, and machine learning, which together form a solid foundation for the development of multi-agent systems [[38]]. These broader efforts are complemented by more specialized research, such as the work at Wake Forest University focusing on the critical challenge of assessing and mitigating the risks associated with multi-agent AI failures—an essential step toward ensuring their safe deployment in real-world scenarios [[39]].

Simultaneously, a highly collaborative research community in Asia, particularly within China, has emerged as an innovation powerhouse. Institutions such as Tsinghua University, Peking University, Zhejiang University, and Shanghai Jiao Tong University are notable for their high volume of AI research and a strong culture of inter-institutional cooperation [[40]]. This collaborative ethos is demonstrated by their prominent and successful participation in academic competitions like the VQualA 2025 Challenge at the International Conference on Computer Vision (ICCV), which highlights their collective capability to solve complex problems that drive the development of advanced agentic systems [[41]]. This networked approach indicates that major breakthroughs are increasingly the product of distributed academic communities. The field's maturation is further evidenced by the prominence of dedicated academic forums. The International Conference on Autonomous Agents and Multiagent Systems (AAMAS) is the discipline's largest and most influential gathering, with its 2025 edition serving as a primary nexus for researchers worldwide [[42]][[43]]. Specialized workshops within AAMAS, like the one on Engineering Multi-Agent Systems (EMAS), signal a growing focus on moving beyond abstract theory to the practical engineering of reliable, scalable, and maintainable agent systems [[44]].

At the leading edge of academic inquiry, researchers are pushing conceptual boundaries with advanced theoretical frameworks and novel architectures. One major direction involves the application of sophisticated game-theoretic models to manage complex agent interactions. The research of Pavel Malinovskiy, for example, extends beyond traditional models to incorporate dynamic coalition formation, the risks of sabotage, and agent behavior in partially observable environments, delivering robust tools for aligning strategic interactions under conditions of uncertainty [[45]]. Another critical frontier is the creation of new architectures for Multi-Agent Reinforcement Learning (MARL). Building on the architectural importance of sequential decision-making discussed previously, the work on the Agent Order of Action Decisions-MAT (AOAD-MAT) model by researchers Shota Takayama and Katsuhide Fujita exemplifies this trend. Their research, which uses a Transformer-based architecture to learn the optimal sequence of agent actions, has achieved significant performance gains on complex benchmarks, pointing toward more efficient coordination mechanisms for AI agent teams [[10]].

Corporate Labs and Commercial Applications

The corporate sector has pivoted aggressively from foundational AI research to the development and deployment of sophisticated multi-agent systems, effectively translating academic theory into practical, high-value tools [[46]]. This shift is motivated by the understanding that coordinated agent teams can achieve unprecedented efficiencies and solve complex problems in both business and science. The industrial landscape is characterized by a dual-pronged approach: large technology firms are leveraging their vast resources to construct general-purpose agentic platforms for grand scientific challenges, while a dynamic startup ecosystem is targeting high-value vertical domains with specialized "AI scientist" solutions [[47]][[48]]. This wave of commercialization is generating a new class of enterprise tools where agents can autonomously operate software, integrate with ERP systems, and execute complex business workflows [[46]][[49]].

Google DeepMind stands as a prime example of the large-scale corporate effort, consistently setting the pace in both fundamental agentic research and its application. Projects like SIMA (Scalable, Instructable, Multiworld Agent) showcase the development of agents capable of learning within complex 3D virtual environments, a foundational skill for interacting with either simulated or physical systems [[50]]. In a direct application of the "AI scientist" paradigm, DeepMind has established its first automated materials science laboratory in the UK. This initiative aims to build a fully autonomous system for discovering and synthesizing novel materials, representing a clear instantiation of a multi-agent system designed for end-to-end scientific research, from hypothesis generation to physical experimentation [[51]].

In parallel with these industrial giants, a vibrant startup scene has emerged to create specialized AI agents for specific scientific and industrial problems. Latent Labs is a prominent leader in computational biology, as evidenced by its 2025 publication on Latent-X, an all-atom protein design model. This AI scientist is engineered for the *de novo* design of protein binders, where it jointly generates both the atomic structure and the amino acid sequence to bind to a specific protein target. The model's unique approach, which directly models non-covalent interactions, has achieved remarkable wet lab validation, demonstrating experimental hit rates over 90% and binding affinities comparable to approved therapeutics. By providing Latent-X through a commercial platform, Latent Labs is not only advancing the state of the art but also commercializing access to highly specialized AI scientists [[52]]. The growth of firms like Latent Labs, alongside others such as Anysphere, Thinking Machine Labs, and World Labs, signals a deepening ecosystem of agentic AI companies [[47]]. This market expansion is further supported by a growing network of AI consulting firms that specialize in integrating these advanced systems into enterprise workflows [[53]][[54]].

The Open-Source Ecosystem: Democratizing Agent Architectures

The rapid innovation and widespread adoption of multi-agent AI are critically supported by a maturing open-source ecosystem. This community serves as an essential bridge, translating theoretical concepts from academia and high-impact applications from corporate labs into accessible, reusable, and extensible frameworks [[55]][[56]]. These open-source projects provide the vital infrastructure that democratizes the creation of sophisticated agentic workflows. The 2025 landscape is defined by a powerful collection of frameworks enabling developers to build everything from simple automated tasks to complex, role-playing teams of AI agents.

Leading open-source projects can be broadly grouped into two categories: orchestration frameworks and comprehensive agent-centric platforms. LangGraph and CrewAI are dominant in the first category. LangGraph, an extension of the popular LangChain library, employs a graph-based paradigm to construct stateful, multi-agent applications. Its architecture is particularly well-suited for modeling the complex, cyclical processes of iteration and reflection inherent in scientific inquiry, allowing developers to represent agent interactions as robust, state-maintaining systems [[57]]. CrewAI complements this by offering a framework specifically designed for orchestrating role-playing, collaborative agents. It simplifies the process of defining agents with specific goals and tools and then assigning them to a team to collectively execute a task [[58]].

In the second category, Microsoft's AutoGen and the community-driven MetaGPT project offer more holistic platforms for building autonomous systems. AutoGen provides a versatile framework for creating applications with multiple, conversable agents that can solve tasks together, supporting both highly deterministic workflows and more dynamic, research-oriented collaborations [[59]]. MetaGPT extends this idea with a highly opinionated, role-based framework that simulates an entire virtual software company. Given just a single-line requirement, MetaGPT can autonomously assign roles (e.g., product manager, architect, engineer) to different agents to generate a complete suite of development documentation, from user stories to API specifications, representing a major step toward specialized, autonomous AI workforces [[60]]. Together, these projects—along with others like Langflow for visual workflow design and Cline for autonomous coding—form a layered ecosystem that is accelerating the entire field [[61]]. By providing standardized protocols and interaction patterns, the open-source movement empowers a global community to build upon foundational research, ensuring that the transformative potential of multi-agent AI is available to the widest possible audience of innovators.

Challenges, Limitations, and Ethical Considerations

As of 2025, the advent of multi-agent AI scientists, exemplified by platforms like Google's AI Co-Scientist and the MASTER framework, signifies a potential paradigm shift in computational research across diverse fields from materials science to drug discovery [[4]][[7]]. However, the path from these promising prototypes to the deployment of robust, autonomous scientific partners is obstructed by a deeply interconnected set of technical challenges, inherent system limitations, and pressing ethical considerations. A critical analysis of these obstacles is essential to navigate the future development and integration of these powerful systems into the scientific enterprise.

Primary Technical Challenges

The successful operation of multi-agent AI scientists hinges on overcoming significant technical hurdles that span individual agent capabilities, collective strategic behavior, and foundational infrastructural capacity. These challenges can be broadly categorized into four primary domains.

First, achieving **strategic coordination and collaborative reasoning** is a foundational obstacle. The objective is to elevate agent teams from simple task parallelism to a state of true synergistic collaboration, a goal complicated by the conditions of partial observability and limited communication common in complex scientific research [[11]]. A critical aspect of this challenge is the temporal sequencing of agent actions. Many multi-agent reinforcement learning (MARL) models do not explicitly account for the strategic importance of action ordering, creating a significant performance gap. The development of specialized architectures like the Agent Order of Action Decisions-MAT (AOAD-MAT), which learns to dynamically determine the optimal sequence of agent actions, underscores the importance of solving this problem [[10]]. Furthermore, existing models are often restricted to purely cooperative scenarios, failing to address the complexities of dynamic coalition formation or the risk of agent sabotage in partially adversarial contexts. This necessitates the integration of more advanced game-theoretic models capable of negotiation and adaptive behavior [[45]].

Second, the ambition of end-to-end autonomous research is constrained by difficulties in **complex planning and workflow orchestration**. A core promise of AI scientists is their capacity to manage multi-step research pipelines, from literature review to experimentation [[62]]. However, a key technical deficit in 2025 is the inadequacy of agents in multi-task planning, where they struggle with multi-goal or multi-tool objectives within a single, coherent workflow [[12]]. This deficit is intrinsically linked to the architectural trade-offs discussed previously. Designers must balance the coordination simplicity of centralized models against their bottleneck risks, while the robustness of decentralized models is complicated by challenges in achieving coherent collective action and mitigating the 'straggler effect' caused by resource heterogeneity [[23]][[24]]. Even sophisticated hybrid architectures like the blackboard system require a complex control shell to manage agent access and facilitate opportunistic problem-solving [[25]]. The success of platforms that orchestrate specialized agents for tasks like 'Literature Analysis' and 'Hypothesis Generation' confirms that workflow management is feasible but also reveals that the design of these orchestration strategies is itself a non-trivial research challenge [[4]].

Third, these systemic challenges are compounded by persistent issues in **grounded cognition and robust scientific reasoning** at the individual agent level. A central goal is to create agents with genuine, domain-specific understanding, such as the "chemically grounded scientific reasoning" demonstrated by the MASTER framework [[7]]. A major impediment to this is ensuring the factual accuracy of agent-generated knowledge. While techniques like Retrieval-Augmented Generation (RAG) are employed to ground agent outputs in established scientific literature, mitigating model hallucination and enabling agents to synthesize coherent insights from vast and sometimes contradictory information remains an ongoing struggle [[16]]. A frontier problem in this area is the translation of high-level human scientific intent into computable agent objectives. The concept of defining agent utilities through natural language, for example, requires a nuanced understanding of scientific goals that is difficult to formalize into traditional reward functions, representing a core challenge in aligning autonomous behavior with human scientific inquiry [[45]].

Finally, the entire endeavor is constrained by the pragmatic challenge of **computational and infrastructural scalability**. The scientific domains where AI scientists show the most promise—such as climate science, quantum materials, and computational chemistry—are characterized by immense search spaces and computationally intensive simulations [[5]][[6]]. The high financial cost and electricity consumption associated with running Density Functional Theory (DFT) calculations or high-resolution Earth System Models on exascale supercomputers act as a significant bottleneck to discovery and place a heavy strain on data center infrastructure [[46]][[6]][[7]]. Consequently, a major research focus is on developing more efficient discovery strategies. The

success of the MASTER framework in reducing the number of expensive simulations by up to 90% exemplifies this priority, highlighting the need to make autonomous discovery computationally and economically viable [[7]].

Principal Limitations of Current Frameworks

Despite clear progress, current multi-agent AI scientist frameworks are defined by profound limitations that underscore the gap between AI as a powerful research *assistant* and the vision of a truly autonomous scientific partner.

- **Emergent Unpredictability and Systemic Brittleness:** The challenges of grounded cognition are magnified in a multi-agent context, where interactions can lead to unpredictable emergent behaviors that cannot be reliably predicted from the constituent agents' individual properties. Research indicates that an ensemble's collective behavior can shift in unexpected ways; for instance, LLM agents have been shown to display 'brittle' moral preferences that change with framing, and agent teams can exhibit emergent group dynamics like 'peer pressure.' These phenomena highlight profound safety and alignment challenges unique to the multi-agent setting [[63]].
- **Incomplete Autonomy and Human-in-the-Loop Necessity:** Even the most celebrated case studies reveal the current boundaries of autonomous operation. Google's AI Co-Scientist excels at literature synthesis and hypothesis generation but ultimately hands off its proposals for human-led validation [[4]][[33]]. The stated goal of the MASTER framework to achieve autonomy *without human intervention in the reasoning loop* implicitly confirms that most contemporary systems still require such intervention [[7]]. This limitation is particularly clear where the digital meets the physical; for example, the AI-driven design of novel protein binders still required subsequent laboratory synthesis and testing to validate the *in-silico* designs, a validation loop that remains dependent on external, non-agentic processes [[13]]. As of 2025, these systems are best characterized as powerful cognitive accelerators, not as independent scientists.

Key Ethical Considerations

The increasing autonomy of AI scientists introduces a complex landscape of ethical issues that demand urgent and proactive consideration from the research community and society at large.

Accountability and Responsibility: A primary ethical challenge lies in assigning responsibility for outcomes generated by systems explicitly designed to minimize human intervention [[1]][[3]]. While current 'co-scientist' models with a human-in-the-loop mitigate this issue by having human experts vet AI-generated outputs [[4]][[33]], the push toward full autonomy raises difficult questions of liability for erroneous, harmful, or unintended results [[7]]. In systems that exhibit emergent, opportunistic reasoning, such as a blackboard architecture, tracing a single causal chain for a flawed output becomes nearly impossible, obscuring whether the fault lies with an individual agent, its architects, or its end-users [[25]].

Dual-Use Risks and Research Security: The unprecedented speed and potential creativity of these platforms introduce significant dual-use risks. A system that can accelerate the design of novel protein binders for cancer therapy in weeks could, in principle, be repurposed to design potent bioweapons with similar efficiency [[13]]. An agent that discovers novel materials for quantum computing could also be directed to design new explosives [[5]]. This acceleration of the design-validate cycle shortens the time available for oversight and intervention. The laudable goal of democratizing access to powerful scientific tools simultaneously lowers the barrier for malicious actors to exploit these capabilities for harmful ends [[27]].

Epistemic Integrity and Reliability: The integration of AI agents into the fabric of science raises fundamental questions about the integrity of the knowledge they produce. Model hallucination remains a core vulnerability of the underlying LLMs, and while RAG offers partial mitigation, it is not a complete solution [[16]]. In a multi-agent system, a single hallucination can be accepted and amplified by other agents, leading to the construction of elaborate but entirely baseless conclusions. This risk has spurred a push for open and transparent AI pipelines and a move away from proprietary, black-box systems [[29]]. However, the inherent opacity of emergent reasoning in systems that learn their own coordination strategies can make it difficult to fully audit their conclusions, challenging core scientific norms of transparency and reproducibility [[10]].

Socio-Economic Equity and Environmental Sustainability: The deployment of AI scientists carries significant socio-economic and environmental implications. Despite their promise to democratize science, the immense financial and computational cost of developing and running these systems risks concentrating the most powerful research tools within a few well-funded labs, thereby exacerbating the global scientific divide. Furthermore, the massive electricity consumption required for the underlying simulations and the AI models themselves presents a profound sustainability challenge [[6]]. This creates an ethical paradox where tools used to solve critical problems like climate change simultaneously contribute to the very resource and energy pressures the world faces. While AI-driven optimizations that reduce simulation counts by up to 90% are a positive step, the

overall energy footprint of this technology necessitates a careful balancing of scientific progress against its environmental cost [[7]].

Future Trajectory: Emerging Trends Beyond 2025

While 2025 has established the value of "Agentic Research" through pioneering systems like Google's AI Co-Scientist and MASTER, these platforms are also defined by their current limitations [[4]][[35]][[7]]. As discussed, they often exhibit incomplete autonomy and brittleness, remaining dependent on a "human-in-the-loop" for critical reasoning and validation [[13]][[33]]. The period beyond 2025 will be defined by a concerted effort to overcome these constraints. The future trajectory points toward greater agent autonomy and more strategic collaboration, the expansion of applications into interdisciplinary grand challenges and "meta-science," and the maturation of the infrastructure, governance, and safety protocols required for responsible innovation.

Evolving Technical Capabilities: Towards Autonomous and Strategic Collaboration

A primary frontier for post-2025 research is the evolution of agent coordination from simple task parallelism to sophisticated, synergistic collaboration. As previously noted, many current multi-agent reinforcement learning (MARL) models do not fully account for the strategic importance of action sequencing and are often limited to purely cooperative scenarios [[10]][[11]]. Future work will aim to transcend these limitations by developing advanced models informed by game theory, enabling agent teams to navigate mixed-motive environments involving negotiation, dynamic coalition formation, and even potential sabotage [[45]]. A key aspect of this will be solving the challenge of translating high-level scientific goals expressed in natural language into machine-interpretable "language-based utilities" [[45]]. Building on the foundation of architectures like AOAD-MAT, which optimizes the temporal dimension of collaboration [[10]], these systems will likely generalize and expand upon the nascent collaborative strategies seen in the MASTER framework, such as 'Peer Review', potentially through meta-learning agents that dynamically invent new collaboration protocols [[7]].

Concurrently, a major objective is to push beyond the prevailing "human-in-the-loop" model toward fully end-to-end discovery cycles [[33]]. While 2025 systems excel at *in-silico* tasks, they still rely on human scientists for physical synthesis and validation [[4]][[13]]. The future vision, actively pursued by labs like Google DeepMind, is the creation of integrated, AI-assisted laboratories where agentic systems design novel materials and directly interface with robotic platforms to autonomously synthesize and test them in a closed loop [[28]][[51]]. Realizing this requires overcoming the previously discussed hurdles in complex workflow orchestration, including the architectural trade-offs between centralized and decentralized models [[23]][[24]][[12]]. Future research will likely focus on robust hybrid architectures, potentially incorporating principles from classical blackboard systems for opportunistic problem-solving, with open-source frameworks like LangGraph providing templates for managing the stateful, cyclical agent interactions characteristic of scientific inquiry [[25]][[57]].

The reliability of these increasingly autonomous systems hinges on deepening their scientific reasoning. As of 2025, genuine, domain-specific understanding—such as the "chemically grounded" reasoning of MASTER—remains an exceptional achievement rather than a standard feature [[7]]. The risk of model hallucination, only partially mitigated by techniques like Retrieval-Augmented Generation (RAG), remains a fundamental threat, as a single fabricated "fact" can be amplified by an agent team [[16]][[63]]. Post-2025 research will pivot towards creating more inherently reliable and self-correcting systems. This will involve not only training models on vast, multi-modal scientific data but also a new research thrust in "AI for AI safety." This entails developing specialized "auditor" or "skeptic" agents whose function is to continuously probe, verify, and challenge the outputs of their peers, creating an internal adversarial process for ensuring factual accuracy and addressing calls for open and verifiable AI pipelines [[29]][[63]].

Expanding Application Frontiers: Grand Challenges and Meta-Science

As technical capabilities mature, the application of multi-agent AI scientists will deepen within existing domains and expand into more complex, interdisciplinary frontiers. Building on 2025 successes in designing protein binders [[13]][[32]], future platforms will aim to manage entire preclinical drug discovery pipelines by integrating genomics, therapeutic design, and clinical trial formulation into holistic, personalized strategies [[12]]. In materials science, agent teams will move beyond single properties to tackle the inverse design of metamaterials. By intelligently managing computationally expensive simulations—as pioneered by

MASTER's 90% reduction in simulation counts—these systems will make exploring astronomically large design spaces feasible, with 'Robotic Synthesis Agents' translating digital blueprints into physical prototypes in automated labs [[28]][[7]][[51]].

Beyond these specializations, the most significant expansion will be into domains characterized by complex, dynamic interactions. Climate science, a nascent area for agentic AI in 2025, may see the development of integrated "digital twins" of the planet that couple Earth System Models (ESMs) with socio-economic models [[6]]. In such a framework, a team of agents representing the atmosphere, oceans, and economic policy would operate in a mixed-motive environment, requiring sophisticated, game-theoretic coordination to simulate feedback loops between climate change and human behavior [[45]]. Agentic systems are also positioned to address scientific "grand challenges" by assembling "virtual institutes" of specialist agents (e.g., 'Microbiology', 'Epidemiology') to tackle problems like antibiotic resistance. Supported by orchestration frameworks like AutoGen or LangGraph, these systems could synthesize insights across disciplinary silos—a major bottleneck in human-led research [[4]][[57]][[59]][[64]][[65]].

Perhaps the most profound future application is "meta-science"—using AI to improve the scientific process itself, directly responding to concerns about epistemic integrity [[29]][[16]]. A high-impact use case will be deploying 'Auditor' agents to perform automated replication and epistemic auditing. These agents could autonomously attempt to replicate peer results, trace data provenance, and design experiments to falsify a team's prevailing hypotheses, institutionalizing skepticism in a continuous loop of automated peer review [[10]][[63]]. The ultimate long-term vision is automated scientific theory generation, where an agent team analyzes decades of data to propose new fundamental principles and derive testable predictions.

Maturing the Ecosystem: Infrastructure, Governance, and Safety

The ambitious trajectory for multi-agent AI scientists is fundamentally linked to the co-evolution of the surrounding ecosystem to address immense computational costs and significant ethical challenges [[46]][[6]][[24]]. On the infrastructure front, scalability will be pursued through algorithmic efficiency, with "meta-optimization" agents inspired by MASTER's success in reducing simulation requirements [[7]]; new compute paradigms; and the maturation of the open-source ecosystem. Projects like AutoGen and CrewAI will likely move toward standardization and interoperability protocols that democratize access, a goal supported by engineering-focused academic workshops like EMAS [[44]][[55]][[58]][[59]].

Simultaneously, governance and safety research must mature to address the ethical considerations of increased autonomy. Research in multi-agent safety, such as that at Wake Forest University, will become a critical subfield, moving beyond single-agent alignment to address the unpredictable emergent behaviors and "brittle" moral preferences observed in agent teams [[39]] [[63]]. A push for "auditable AI" will demand platforms that provide transparent, reproducible logs of an agent team's reasoning process [[10]]. To mitigate the significant dual-use risks previously identified—where a system for designing therapeutic proteins could also design bioweapons [[13]]—new governance models may include digital watermarks in AI-generated outputs and regulatory frameworks for high-risk autonomous research, especially as democratization lowers the barrier for malicious actors [[27]]. Finally, a concerted effort will be required to balance the immense costs of development with the goal of equitable access, addressing both the risk of a widening scientific divide and the environmental paradox of using energy-intensive AI to solve critical problems like climate change [[6]].

References

[[1]] What Is Multi-Agent AI? Definition, Benefits, and Examples - https://www.newhorizons.com/resources/blog/multi-agent-ai

[[2]] What is a Multi-Agent System? - https://www.ibm.com/think/topics/multiagent-system

[[3]] What is a multi-agent system in AI? - https://cloud.google.com/discover/what-is-a-multi-agent-system

[[4]] Accelerating scientific breakthroughs with an AI co-scientist - https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/

[[5]] A Multi-Agent Framework for Assisting Quantum Materials … - https://cic.ubc.ca/project/a-multi-agent-framework-for-assisting-quantum-materials-research/

[[6]] Climate model - https://en.wikipedia.org/wiki/Climate_model

[[7]] Hierarchical Multi-agent Large Language Model Reasoning for Autonomous Functional Materials Discovery - https://arxiv.org/pdf/2512.13930v1

[[8]] Real-World Applications of AI Agents in Scientific … - https://medium.com/tech-ai-made-easy/real-world-applications-of-ai-agents-in-scientific-research-2025-f741fa65714d

[[9]] Agentic Research: The Future Scientist's Workspace - https://www.causaly.com/blog/ai-agent-platform-for-scientists

[[10]] AOAD-MAT: Transformer-based multi-agent deep reinforcement learning model considering agents' order of action decisions - https://arxiv.org/pdf/2510.13343v1

[[11]] Augmenting the action space with conventions to improve multi-agent cooperation in Hanabi - https://arxiv.org/pdf/2412.06333v3

[[12]] Risks of AI scientists: prioritizing safeguarding over autonomy - https://www.nature.com/articles/s41467-025-63913-1

[[13]] 2025 in science - https://en.wikipedia.org/wiki/2025_in_science

[[14]] AI-Researcher: Autonomous Scientific Innovation - https://arxiv.org/pdf/2505.18705v1

[[15]] AI Agents to MultiAgent Systems: A Capability Framework - https://kenhuangus.medium.com/ai-agents-to-multiagent-systems-a-capability-framework-23836e7dda07

[[16]] Foundations of GenIR - https://arxiv.org/pdf/2501.02842v1

[[17]] Agentic AI for Scientific Research: Autonomous … - https://www.sapiosciences.com/blog/agentic-ai-for-scientific-research-autonomous-agents-transforming-experiment-design/

[[18]] Automating RT Planning at Scale: High Quality Data For AI Training - https://arxiv.org/pdf/2501.11803v5

[[19]] From Models to Scientists: Building AI Agents for Scientific … - https://kempnerinstitute.harvard.edu/research/deeper-learning/from-models-to-scientists-building-ai-agents-for-scientific-discovery/

[[20]] AI Agents Capabilities and Risks: What You Must Know Now - https://www.lumenova.ai/blog/ai-agents-capabilities-risks/

[[21]] Everything you need to know about multi AI agents in 2025 - https://springsapps.com/knowledge/everything-you-need-to-know-about-multi-ai-agents-in-2024-explanation-examples-and-challenges

[[22]] LLM-Based Multi-Agent Systems - https://www.emergentmind.com/topics/llm-based-multi-agent-systems-c730e815-6eb0-4219-a652-47653955a70d

[[23]] LLM-based Multi-Agent Blackboard System for Information … - https://arxiv.org/html/2510.01285v1

[[24]] Communication-Efficient Training Workload Balancing for Decentralized Multi-Agent Learning - https://arxiv.org/pdf/2405.00839v1

[[25]] Blackboard system - https://en.wikipedia.org/wiki/Blackboard_system

[[26]] Learning Bilateral Team Formation in Cooperative Multi-Agent Reinforcement Learning - https://arxiv.org/pdf/2506.20039v1

[[27]] AI agents set to democratise computational chemistry - https://www.chemistryworld.com/news/ai-agents-set-to-democratise-computational-chemistry/4022465.article

[[28]] AI materials discovery now needs to move into the real world - https://www.technologyreview.com/2025/12/15/1129210/ai-materials-science-discovery-startups-investment/

[[29]] Biomedical Informatics and Data Science - https://www.uab.edu/medicine/news/informatics/uab-sparc-showcases-breakthrough-in-ai-agent-driven-drug-discovery-at-global-ai-science-conference

[[30]] Leading artificial intelligence–driven drug discovery platforms - https://www.sciencedirect.com/science/article/abs/pii/S0031699725075118

[[31]] 6 ways AI reshaped scientific software in 2025 - https://www.rdworldonline.com/6-ways-ai-reshaped-scientific-software-in-2025/

[[32]] Google's AI co-scientist just solved a biological mystery that … - https://www.psypost.org/googles-ai-co-scientist-just-solved-a-biological-mystery-that-took-humans-a-decade/

[[33]] Google co-scientist can crunch early hypothesis generation … - https://www.rdworldonline.com/google-ai-co-scientist-can-reduce-early-hypothesis-generation-from-weeks-to-days-in-some-cases/

[[34]] Artificial Intelligence and Generative Models for Materials Discovery -- A Review - https://arxiv.org/pdf/2508.03278v1

[[35]] The Year of the AI Agent in Higher Education - https://www.higher-education-marketing.com/blog/the-year-of-the-ai-agent-in-higher-education

[[36]] Predictions for AI in 2025: Collaborative Agents, AI Skepticism … - https://hai.stanford.edu/news/predictions-ai-2025-collaborative-agents-ai-skepticism-and-new-risks

[[37]] The 2025 AI Index Report | Stanford HAI - https://hai.stanford.edu/ai-index/2025-ai-index-report

[[38]] 9 Best Universities For Artificial Intelligence Reserach in … - https://theaiinsider.tech/2024/05/15/top-9-universities-for-artificial-intelligence-research/

[[39]] Multi-agent AI could change everything – if researchers can … - https://news.wfu.edu/2025/12/22/multi-agent-ai-could-change-everything-if-researchers-can-figure-out-the-risks/

[[40]] Top 15 Pioneering AI Research Institutions : Companies, … - https://medium.com/@joycebirkins/top-15-pioneering-ai-research-institutions-across-china-and-the-us-companies-labs-and-f07f5a495b63

[[41]] VQualA 2025 Challenge on Engagement Prediction for Short Videos: Methods and Results - https://arxiv.org/pdf/2509.02969v1

[[42]] AAMAS 2025 Detroit - https://aamas2025.org/

[[43]] AAMAS: International Conference on Autonomous Agents … - https://dl.acm.org/doi/proceedings/10.5555/3709347

[[44]] Workshops – AAMAS 2025 Detroit - https://aamas2025.org/index.php/conference/program/accepted-workshops/

[[45]] Advanced Game-Theoretic Frameworks for Multi-Agent AI Challenges: A 2025 Outlook - https://arxiv.org/pdf/2506.17348v1

[[46]] AI agents arrived in 2025 – here's what happened and … - https://theconversation.com/ai-agents-arrived-in-2025-heres-what-happened-and-the-challenges-ahead-in-2026-272325

[[47]] Forbes 2025 AI 50 List - Top Artificial Intelligence … - https://www.forbes.com/lists/ai50/

[[48]] AI 100: The most promising artificial intelligence startups of … - https://multiversecomputing.com/resources/ai-100-the-most-promising-artificial-intelligence-startups-of-2025

[[49]] 10 Real-World Examples of AI Agents in 2025 - https://www.xcubelabs.com/blog/10-real-world-examples-of-ai-agents-in-2025/

[[50]] Research - https://deepmind.google/research/

[[51]] Google DeepMind & The UK: The First Automated … - https://aimagazine.com/news/google-deepmind-the-uk-the-first-automated-ai-science-lab

[[52]] Latent-X: An Atom-level Frontier Model for De Novo Protein Binder Design - https://arxiv.org/pdf/2507.19375v1

[[53]] Top 25 AI Consultants & Leaders of 2025 - https://www.theconsultingreport.com/the-top-25-artificial-intelligence-consultants-and-leaders-of-2025/

[[54]] 15 Best Agentic AI Companies of 2025 - https://wotnot.io/blog/best-agentic-ai-companies

[[55]] The 4 Best Open-Source Multi-Agent AI Frameworks in 2025 - https://medium.com/coding-nexus/the-4-best-open-source-multi-agent-ai-frameworks-in-2025-81e92f23f866

[[56]] Top 11 Open-Source Autonomous Agents & Frameworks in … - https://cline.bot/blog/top-11-open-source-autonomous-agents-frameworks-in-2025

[[57]] LangGraph - https://www.langchain.com/langgraph

[[58]] Open source - https://www.crewai.com/open-source

[[59]] AutoGen - https://microsoft.github.io/autogen/stable//index.html

[[60]] FoundationAgents/MetaGPT: ✨ The Multi-Agent Framework - https://github.com/FoundationAgents/MetaGPT

[[61]] Top 9 AI Agent Frameworks as of December 2025 - https://www.shakudo.io/blog/top-9-ai-agent-frameworks

[[62]] How AI agents will change research: a scientist's guide - https://www.nature.com/articles/d41586-025-03246-7

[[63]] MAEBE: Multi-Agent Emergent Behavior Framework - https://arxiv.org/pdf/2506.03053v2

[[64]] The Future of Innovation is Interdisciplinary and AI-Enhanced - https://abramanders.substack.com/p/the-future-of-innovation-is-interdisciplinary

[[65]] The rise of multidisciplinary research stimulated by AI - https://www.insidehighered.com/opinion/columns/online-trending-now/2025/01/22/rise-multidisciplinary-research-stimulated-ai