THE HITCHHIKER'S GUIDE TO AUTONOMOUS RESEARCH: A SURVEY OF SCIENTIFIC AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The advancement of LLM-based agents is redefining AI for Science (AI4S) by enabling autonomous scientific research. Prominent LLMs exhibited expertise across multiple domains, catalysing constructions of domain-specialised scientific agents. Nevertheless, the profound epistemic and methodological gaps between AI and the natural sciences still impede the systematic design, training, and validation of these agents. This survey bridges the existing gap by presenting an exhaustive blueprint for scientific agents, spanning systematic construction methodologies, targeted capability enhancement, and rigorous evaluations. Anchored in the canonical scientific workflow, this paper (i) pinpoints the overview of scientific agents, starting with the development from general-purpose agents to scientific agents driven by articulated goal-orientation, then subsequently advancing a comprehensive taxonomy that organises existing agents by construction strategy and capability scope, and (ii) introduces a two-tier progressive framework, from scientific agents contrustion from scratch to targeted capability enhancement, for realizing autonomous scientific research. It is our aspiration that this survey will serve as guidance for researchers across various domains, facilitating the systematic design of domain-specific scientific agents and stimulating further innovation in AI-driven scientific research. To support long-term progress, we curate a live repository (AWESOME_SCIENTIFIC_AGENT) that continuously aggregates emerging methods, benchmarks, and best practices.

1 INTRODUCTION

The scientific research paradigm is progressively transitioning from a computational paradigm to the fifth paradigm, known as scientific intelligence Bishop (2022), with the swift advancement of artificial intelligence technology, particularly in the domain of Large Language Models (LLMs) OpenAI (2022); Anthropic (2024); Gemini Team, Google DeepMind (2023); DeepSeek-AI et al. (2024); Bai et al. (2023).

Despite the continuous transformation of paradigms in scientific research, the basic lifecycle of research process remains largely unchanged, typically comprising six stages as shown in Fig. 1: literature mining, research hypothesis, experiment design, experiment verification, analysis and result, and finally, evaluation and review. Former scientific research paradigms are frequently predicated on varying research subjects. The primary impetus for innovation in experimental design and the verification process continues to arise from the knowledge framework of scientific researchers. With the advent of LLMs that exhibit capabilities in natural language understanding and deductive reasoning, artificial intelligence (AI) can engage in the scientific research process at the granularity of human knowledge, rather than serving merely as a tool for a particular procedural element Xu et al. (2021); Boiko et al. (2023a). Concurrently, due to advancements in existing AI agents concerning tool integration and task collaboration, AI can execute specific task flows from start to finish, thus enabling LLM-based scientific agents to support or potentially play a role in any process within the scientific research lifecycle.

Scientific agents, propelled by leading enterprises, have already attained initial success in probing the boundaries of science, as exemplified by Alphaevolve Novikov et al. (2025) in algorithm innovation and Robin Ghareeb et al. (2025) in drug discovery, suggesting considerable potential. While it is posited that extant LLM-based scientific agents have scientific research possesse adequate capacity

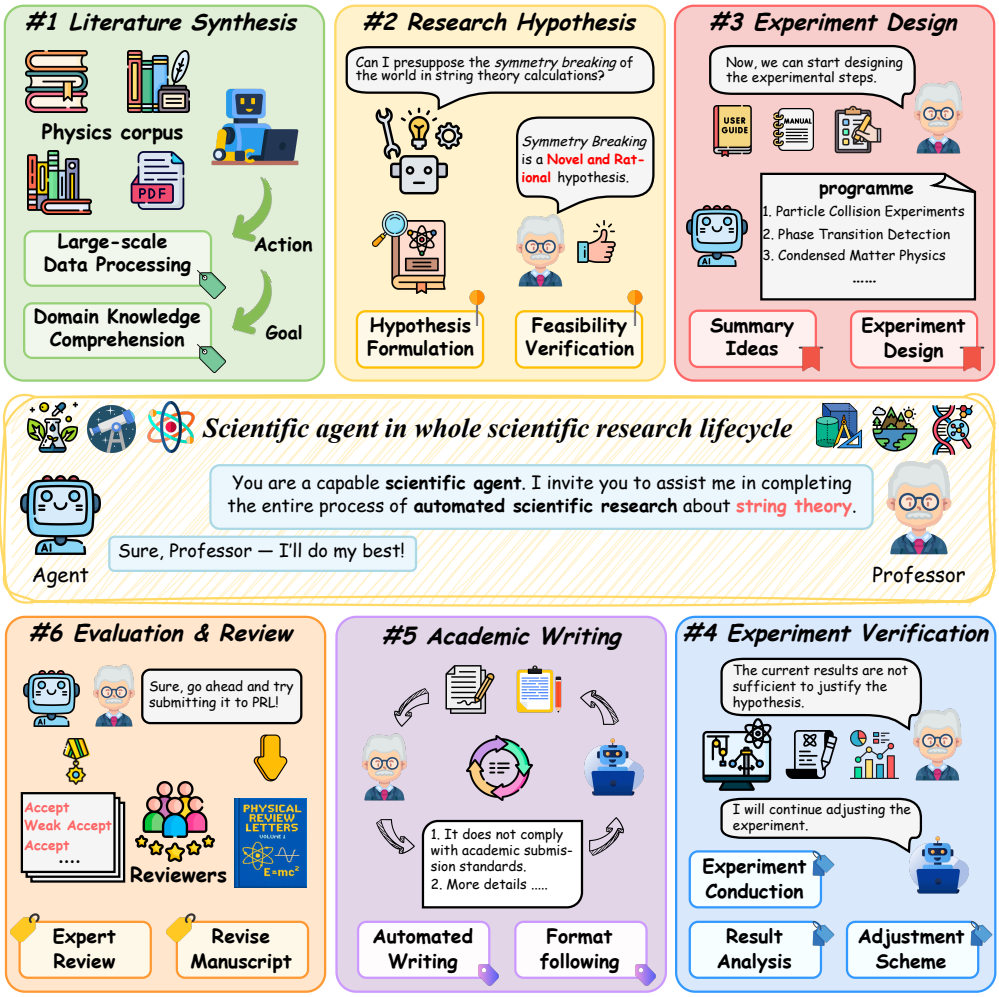


Figure 1: The whole scientific research lifecycle with scientific agents. We use the example of super-asymmetry in *The Big Bang Theory* to explain the process.

to unearth novel scientific findings, LLMs frequently require specialization and tuning for sub-fields within specific disciplines. This necessity partially restricts their seamless integration with particular scientific challenges. The interdisciplinary character of AI for Science further exacerbates this divide across fields. AI researchers Wang et al. (2023b) often lack a nuanced understanding of specialized domain data, while domain experts may not be well-versed in the latest advancements in AI technology. This discrepancy has somewhat impeded the progression of scientific intelligence. Despite the burgeoning emergence of research agents in diverse incarnations, there is an absence of systematic guidance and a unified direction.

Table 1: Comparison of recent surveys of LLM-based scientific research.

Study	Key Focus	Field Analysis					
		Agent Development	Agent Taxonomy	Research Process	Construction	Capability Statement	Evaluations
Gridach et al.Gridach et al. (2025)	Agentic AI development in Science.	✓	✗	✗	✗	✗	✓
Ren et al. Ren et al. (2025a)	Research agent structure decomposition.	✓	✓	✗	✗	✗	✓
Zheng et al. Zheng et al. (2025)	Hierarchy of LLM-based scientific research methodologies.	✗	✓	✗	✗	✗	✓
Chen et al. Chen et al. (2025b)	Application of AI in different research processes	✓	✓	✓	✗	✗	✓
Ours	Construction of scientific agents with various capabilities.	✓	✓	✓	✓	✓	✓

As shown in Table 1, several surveys examine LLMs in scientific research. Gridach et al. Gridach et al. (2025) explored the impact of Agentic AI, while Ren et al. Ren et al. (2025a) detailed research agent framework components. Zheng et al. Zheng et al. (2025) categorized methodologies by their research application efficacy, and Chen et al. Chen et al. (2025b) reviewed the AI-driven research process. However, these surveys do not cover the methodology for constructing scientific agents and the discussion of their capabilities. This review aims to fill these gaps.

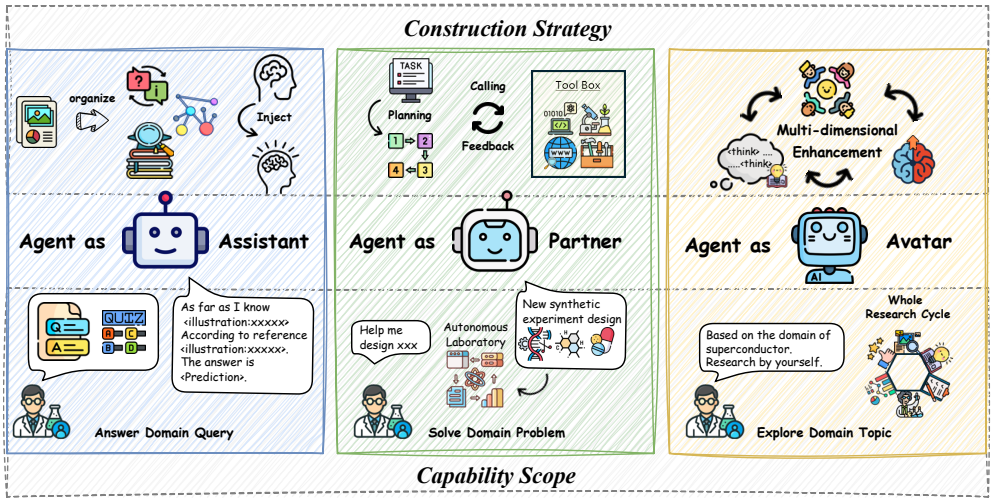


Figure 2: Scientific agents from different levels vary on construction strategy and capability scope. From left to right, respectively, show the focus of the corresponding level of agents (Assistant, Partner, Avatar).

he principal contributions of this survey are as follows:

- While encompassing a broad spectrum of scientific agents, we specifically concentrate on a thorough and rigorous deconstruction of scientific agents within the natural sciences.
- We provide comprehensive and fine-grained practical guidance on the foundational processes involved in constructing scientific agents from scratch, alongside advanced strategies for enhancing the capabilities of existing agents.
- By integrating the scientific research life cycle with a cohesive construction strategy, we propose a novel connection between the design and application of scientific agents, which has yet to be explored in prior literature.

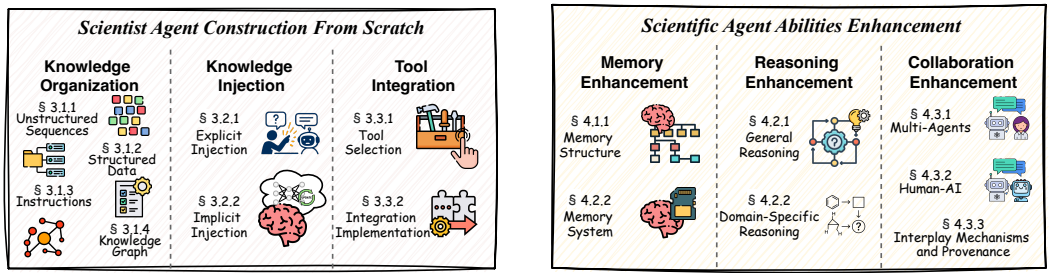
This survey aspires to serve as the hitchhiker's guide to autonomous research with guidance on the design and implementation of existing research agents, thereby facilitating the convergence of AI research and natural scientific research.

2 Overview of Scientific Agents

2.1 From General-purpose Agent to Scientific Agent

With the ongoing advancements in artificial intelligence, the definition of agent has progressed into that of a closed-loop system which perceives its environment, executes actions, and enhances itself through feedback mechanisms Luo et al. (2025). The rapid expansion of LLMs has enlarged the applications of agents, ranging from personal assistants and game AI to agentic AI Sapkota et al. (2025) and scientific research.

As Shohams seminal delineation of an agent's goal-directedness Shoham (1993), the intrinsic nature of objectives profoundly influences its architectural design and operational methodologies. General-purpose agents pursue a **utility-oriented** logic: they prioritize adaptability across multiple scenarios and an improved user experience, addressing diverse requirements through real-time interaction and algorithmic optimization. Conversely, scientific agents are motivated by an inherent **truth-seeking** logic, oriented towards scientific inquiry, theory formulation, and hypothesis assessment. These goals necessitate more stringent criteria on knowledge organization, conceptual reasoning, and protocol integration, seamlessly with experiments.



(a) Agent construction methodology.

(b) Agent enhancement methodology.

Figure 3: Overview of (a) agent construction methodology and (b) Agent enhancement methodology.

2.2 Taxonomy For Scientific Agent

The intersection of LLMs with domain-specific scientific challenges has generated an expanding scope of research. By analyzing the prominent inductive features of representative studies, we introduce a three-tier taxonomy that encapsulates the progressive construction strategies and capability scope in Fig. 2. Due to space limitations, we have selected representative methods and listed them in Table 2. Additional construction strategies and delineations of capability scopes for each level of research agent are presented in Appendix A.

3 Scientific Agent Construction from Scratch

For a specific domain, how to construct a scientific agent from scratch must be solved before autonomous research, as shown in Fig. 3(a). This section mainly clarifies the intelligent agent construction pathway from three aspects: knowledge organization, knowledge injection, and tool integration.

3.1 Knowledge Organization

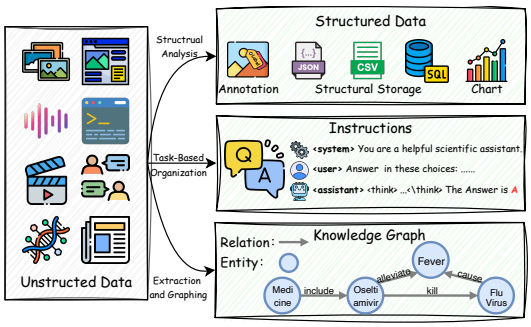


Figure 4: Knowledge organization in scientific agents.

Knowledge organization fundamentally conditions an agents capacity to comprehend existing information and drive scientific discovery: scientific knowledge comprises discrete units whose arrangement governs reasoning and generalization, and varying sparsitydensity regimes reflect distinct information-processing and cognitive models. We survey four paradigms (Fig. 4) unstructured sequence, structured data, instruct input, and knowledge grapheach specifying how knowledge is represented, accessed, and exploited; together they underpin perception, inference, and decision-making in scientific agents.

3.1.1 Unstructured Sequence

Unstructured sequences-such as books, research papers, and lab reports-constitute the primary format in which scientific knowledge is represented in the real world. These rich professional corpora provide strong priors that empower scientific agents to effectively support automated scientific research. As humans do, scientific agents also rely on publicly available articles and preprints across disciplines to acquire domain-specific knowledge from complex unstructured corpora Schmidgall et al. (2025); Cappello et al. (2025); Li et al. (2025c); Lu et al. (2024); Huang et al. (2024c); Agarwal et al. (2024). Specifically, in the biomedical domain, open-access literature databases such as PubMed, PubMed Central (PMC), and PubChem serve as valuable resources for direct knowledge retrieval. Besides, for programming and code-related research, open-source repositories and

Table 2: Taxonomy for scientific agents. Components: **T** indicates *Tool Integration*, **R** indicates *Reasoning Enhancement*, **M** indicates *Memory Enhancement*, and **C** indicates *Collaboration enhancement*. Application Stages is the research process stage in which the method is involved: (L) indicates *Literature Mining*, (H) indicates *Research Hypothesis*, (D) indicates *Experiment Design*, (V) indicates *Experiment Verification*, (A) indicates *Analyst and Result*, (E) indicates *Evaluation and Review*.

Level	Methods	Domains	Base Model	Components				Application Stages	Tasks Descriptions
				T	R	M	C		
Agent As Assistant	AstroLLaMA-Chat Nguyen et al. (2024)	Astronomy	LLaMA	✗	✗	✗	✗	D A	Astronamy knowledge Q&A
	BioGPT Luo et al. (2023)	Biology	GPT-2	✗	✗	✗	✗	L A E	Domain knowledge Q&A and relation extraction
	DARWIN 1.5 Xie et al. (2024)	Biology& Chemistry	LLaMA-7B	✗	✗	✗	✗	L D A	Domain knowledge Q&A
	ChemBERTa Chithra et al. (2020)	Chemistry	RoBERTa	✗	✗	✗	✗	L A E	Entity & Relation extraction and text classifaction
	ChemAU Liu et al. (2025d)	Chemistry	Qwen-Series & LLaMA3-8B	✗	✓	✗	✓	L A	Domain Knowledge Q&A
	ChemDFM Zhao et al. (2024b)	Chemistry	LLaMa-13B	✗	✗	✗	✗	L A	Molecules recognition & property prediction & design
	LlaSMol Yu et al. (2024)	Chemistry	LLaMA-2 & Mistral	✗	✓	✗	✗	D A E	Molecules description & property prediction and reaction design
	InstructMol Cao et al. (2023)	Chemistry	Vicuna-7B	✗	✗	✗	✗	D A E	Molecule description and reaction analysis
	Ether0 Narayanan et al. (2025a)	Chemistry	Mistral-24B	✗	✓	✗	✗	D V	Complex molecule design with reasoning
	PredictiveChem Jablonka et al. (2024)	Chemistry	GPT-3	✗	✗	✗	✗	L D A	Chemical property classification & regression and inverse design
	MoleculeSTM Liu et al. (2023b)	Chemistry	BERT & GraphMVP	✗	✗	✗	✗	A E	Text retrieval and molecule editing
	ClimateGPT Thulke et al. (2024)	Climate	LLaMA-2	✓	✗	✗	✗	L A E	Domain knowledge Q&A
	HypoGen O'Neill et al. (2025)	Computer Science	LLaMA-3.1-8B	✓	✓	✗	✗	L H E	Hypothesis generation
	DeepSeek-Prover-V2 Ren et al. (2025b)	Mathematics	DeepSeek-Prover-V2-7B	✗	✓	✗	✗	V E	Formal proof generation
	BiMediX Pieri et al. (2024)	Medical	Mixtral-8x7B	✗	✗	✗	✗	D A E	Doctor-patient consultation simulation
	ChatDoctor Li et al. (2023)	Medical	LLaMA-7B	✓	✗	✗	✗	D A E	Doctor-patient consultation simulation
	AgentMD Jin et al. (2024a)	Medical	GPT-Series	✓	✗	✗	✓	D A E	LLM-enhanced clinical analysis
	MedAlpaca Han et al. (2023)	Medical	LLaMA	✗	✗	✗	✗	D A E	Medical scene Q&A
	DrugGen Sheikholeslami et al. (2025)	Medical	GPT-2	✗	✗	✗	✗	D V	Molecule design & generation
	LLM-SR Shojaee et al. (2025a)	Physics	GPT-3.5 & Mixtral-8x7B	✗	✓	✓	✗	D A E	Equation Discovery
	LitLLM Agarwal et al. (2024)	General	GPT-Series	✓	✗	✗	✗	L	Scientific literature review generation
	SciBERT Beltagy et al. (2019)	General	BERT-Base	✗	✗	✗	✗	V A E	Entity & Relation extraction and text classifaction
	SciMON Wang et al. (2023c)	General	GPT/ T5	✓	✓	✗	✗	L H E	Science Hypothesis Generation
	SciTune Horawalavithana et al. (2023)	General	LLaVa	✗	✗	✗	✗	L A	Science chart Q&A
	NatureLM Xia et al. (2025)	Hybrid	Mixtral-8x7B	✗	✓	✗	✗	D A	Multimodal-sequence perception and design
Agent As Partner	StarWhisper Wang et al. (2025a)	Astronomy	Qwen-2.5	✓	✓	✓	✓	L D V A E	Telescope control workflow based on LLM agent
	BioResearcher Luo et al. (2024)	Biomedical	GPT-Series	✓	✓	✗	✗	L D V A E	Biological experiment design and implementation
	Crispr-GPT Huang et al. (2024a)	Biology	GPT-Series	✓	✓	✓	✗	L D V A E	LLM-based Agent assists Crispr experiment
	DrBioRight 2.0 Liu et al. (2025c)	Biology	LlaMA-3 & ChatGPT	✓	✓	✗	✗	L D V A E	LLM-assisted bioinformatics platform
	ProtAgents Liu et al. (2024b)	Biology	GPT-4	✓	✓	✗	✓	L D V A E	Protein design & Analysis
	MOOSE-Chem Yang et al. (2024b)	Chemistry	GPT-4o	✓	✗	✓	✗	L H E	Chemistry Hypothesis Generation
	Coscientist Boiko et al. (2023b)	Chemistry	GPT-4	✓	✓	✗	✗	L D V A	Chemical experiment design implementation
	ChemCrow Bran et al. (2023)	Chemistry	GPT-4	✓	✓	✗	✗	L D V A E	Chemical experiment design implementation
	Organa Darvish et al. (2025)	Chemistry	GPT-3.5	✓	✗	✗	✗	V A	Chemical experiment implementation with robot
	xChemAgents Polat et al. (2025)	Chemistry	LLaMA-3	✓	✓	✗	✓	L D V A E	Molecular properties prediction & interpretation
	ChemAgent Tang et al. (2025b)	Chemistry	GPT-Series& Qwen-2.5-72b	✓	✓	✓	✗	L D A	Complex chemistry reasoning question
	Chemma Zhang et al. (2025d)	Chemistry	LLaMA-2-7B	✓	✓	✗	✗	D A	property prediction and synthetic experiment design
	MyCrunchGPT Kumar et al. (2023)	Physics	GPT-3	✓	✗	✗	✗	D V A	LLMs assist SciML workflow design
	Meta-OpenFoam Chen et al. (2024a)	Physics	GPT	✓	✓	✗	✓	D V A E	CFD Simulation Completed
	FoamAgent Yue et al. (2025)	Physics	Claude-3.5-Sonnet	✓	✓	✗	✓	D V A E	CFD Simulation Completed
	AI-Researcher Tang et al. (2025a)	Computer Science	Claude & GPT	✓	✓	✗	✓	L H D V A E	Autonomous scientific exploration in computer science
	Jupybara Wang et al. (2025b)	Computer Science	GPT-4	✓	✓	✗	✓	D V A E	Complex data mining interactions
	FlowAgent Shi et al. (2025)	Computer Science	General LLM	✓	✓	✗	✗	D V A E	Workflow Extensions
	AI Scientist Lu et al. (2024)	Computer Science	GPT & Claude	✓	✗	✗	✗	L H D V A E	Autonomous scientific exploration in computer science
	GeoGPT Zhang et al. (2023a)	Geopgrahy	GPT-3.5	✓	✗	✓	✗	D V A	LLMs use GIS to complete geospatial tasks
	PaperQA Lála et al. (2023)	General	GPT-Series	✓	✓	✓	✗	L A E	Answering questions from scientific documents
	ChatCite Li et al. (2024f)	General	GPT-3.5 & GPT-4	✓	✓	✓	✗	L A	Literature digest & assesment
	PiFlow Pu et al. (2025)	General	GPT-Series	✓	✓	✗	✓	H D V A	Experiment design & verification by scientific principles
	Aviary Narayanan et al. (2025b)	Hybrid	LLaMA-3.1-8B-Instruct	✓	✓	✗	✗	L D V A E	Challenging scientific tasks on small LLMs
	A-Lab Szymanski et al. (2023)	Materials	N/A	✓	✗	✓	✗	D V A E	Materials synthesis recipe & experiment excution
	KG-FM Bai et al. (2025)	Materials	Qwen2-72B	✓	✓	✗	✗	L H A	KG-based materials knowledge Q&A
	Multicrossmodal Bazgir et al. (2025)	Materials	Gemini-2.0 & DeepSeek-R1	✓	✗	✗	✓	L A E	Multimodal materials knowledge Q&A
	LLMatDesign Jia et al. (2024)	Material	GPT-4o	✓	✗	✗	✗	D A E	Material design for specific target properties
	HoneyComb Zhang et al. (2025b)	Material	GPT&LLaMA	✓	✓	✗	✓	D V A E	Material experiment design & impletation
	DrugAgent Liu et al. (2025b)	Medical	Claude & GPT	✓	✓	✗	✓	V D V A E	Pharmacokinetic properties prediction and HTS assistant
	TAIS Liu et al. (2024a)	Medical	N/A	✓	✓	✗	✓	D V A	Gene expression analysis
	MedAgents Tang et al. (2024)	Medical	GPT-4	✓	✓	✗	✓	L A E	Medical Reasoning
	otto-SR Cao et al. (2025a)	Medical	GPT-4	✓	✓	✗	✗	L A E	In-depth Systematic reviews
	MRAgent Xu et al. (2025b)	Medical	GPT	✓	✓	✗	✗	L V A E	Use Mendelian Randomization enhance medical inference
	GeneGPT Jin et al. (2024b)	Medical	CodeX	✓	✓	✗	✗	V A E	Biomedical Reasoning Q&A
Agent As Avatar	CycleResearcher Weng et al. (2024)	Computer Science	Mistral & Qwen-2.5	✓	✓	✗	✓	L H D V A E	Autonomous scientific exploration in computer science
	AI Scientist-v2 Yamada et al. (2025)	Computer Science	GPT & Claude	✓	✓	✗	✗	L H D V A E	Autonomous scientific exploration in computer science
	Agentrxiv Schmidgall & Moor (2025)	Computer Science	GPT-Series	✓	✓	✓	✓	L H D V A E	Autonomous scientific exploration in computer science
	Agent Laboratory Schmidgall et al. (2025)	Computer Science	GPT-Series	✓	✓	✓	✓	L H D V A E	Autonomous scientific exploration in computer science
	Robin Ghareeb et al. (2025)	Biology	GPT & Claude & Gemini	✓	✓	✓	✓	L H D V A E	Autonomous scientific exploration
	AlphaEvolve Novikov et al. (2025)	General	Gemini-Series	✓	✓	✓	✗	L H D V A E	Autonomous scientific exploration
	Biomni Huang et al. (2025a)	Biology	GPT-Series	✓	✓	✓	✓	L H D V A E	Autonomous scientific exploration with wide range of experiments
	OriGene Zhang et al. (2025e)	Medical	Gemini-Series	✓	✓	✓	✓	L H D V A E	Scientific plan formulation and results analysis
	CellVoyager Alber et al. (2025)	Biology	GPT-4o	✓	✓	✗	✗	L H D V A E	ScRNA-seq autonomous exploration
	Sparks Ghafarollahi & Buehler (2025c)	Biology	GPT-4	✓	✓	✓	✓	L H D V A E	Autonomous scientific exploration
	AI co-scientist Gottweis et al. (2025)	Biology	Gemini-2.0	✓	✓	✓	✓	L H D V A E	The entire scientific research process

datasets on platforms like GitHub and Hugging Face complement academic papers by providing practical implementation knowledge. In summary, the large-scale, diverse, and information-rich nature of unstructured sequences makes them indispensable for enabling scientific agents to conduct literature understanding, hypothesis generation, and domain-specific reasoning.

3.1.2 Structured Data

Compared to unstructured corpora, structured data offers denser and more fine-grained knowledge, enabling agents to tackle more complex scientific tasks. Such structured information can be extracted and organized from unstructured sequences or other non-textual domain knowledge forms, such as material crystal structures, chemical formulas of drugs, and so on. In numerous studies, certain methods directly utilize existing structured data from their respective domains to facilitate scientific discovery. For example, TAIS Liu et al. (2024a) leverages structured gene expression datasets to train models capable of identifying predictive genes associated with specific diseases. StarWhis-

per Wang et al. (2025a) integrates telescope observations, weather data, and historical records to enable automated and adaptive astronomical observation.

3.1.3 INSTRUCTIONS

As the most comprehensive form of knowledge organization, instructions are categorized into instruction context and instruction pairs according to their subsequent modes of knowledge injection. Instruction context facilitates prompt engineering for effective knowledge incorporation by partitioning prompts into `<system>` and `<user>` inputs that specify requirements and an `<assistant>` output, with user-provided exemplars functioning as few-shot context to improve intuitiveness, followability, and response standardization (Fig. 4); state-of-the-art agents such as Agentrxiv Schmidgall & Moor (2025), Robin Ghareeb et al. (2025), and Biomni Huang et al. (2025a) employ this paradigm to steer complex tasks. Complementarily, instruction pairs formalize data as ordered questionanswer tuples that both inject scientific knowledge and strengthen instruction following: Xu et al. Xu et al. (2025a) draw on MMLU-pro Wang et al. (2024c) across Math, Health, Business, and Law to study collaborative expertise specialization; SciTune Horawalavithana et al. (2023) leverages SciCap Hsu et al. (2021) and ScienceQA Lu et al. (2022) in a two-stage pipeline (scientific concept alignment scientific instruction fine-tuning); and ClimateGPT Thulke et al. (2024) curates expert-grounded, domain-specific instruction datasets via interviews and demonstrations.

3.1.4 KNOWLEDGE GRAPH

Knowledge graphs (KGs) encode scientific knowledge as triples, enabling semantic reasoning, explainability, efficient retrieval, and tight integration with RAG for discovery. SciMON integrates three KG signalssemantic-similarity graphs for nearest-neighbor retrieval of prior work, global KG neighbors for related concepts, and citation networks for literature neighborsto ground hypothesis generation Wang et al. (2023c). Similarly, the Auto Research framework injects KG-structured knowledge across the workflow, using graph-based mapping/categorization of literature to surface research gaps and promising directions Liu et al. (2025a).

3.2 KNOWLEDGE INJECTION

Knowledge injection can be categorized into two types: explicit and implicit. The former requires no modification of model parameters, while the latter typically integrates external knowledge by further fine-tuning LLMs or incorporating lightweight modules. Table 3 lists representative frameworks related to knowledge injection, which significantly lowers the technical barrier to entry for domain knowledge injection and simplifies the entire process.

Table 3: Mainstream frameworks for knowledge injection.

Method	Applications	Features	Preprations Input	Target Output
PromptAgent Wang et al. (2024a)	Prompt Optimization	MCTS optimization process	Raw prompt & Domain Task Input	Optimized prompts
Promptfoo community (2025)	Prompt Optimization	Testdriven evaluation loop	Raw Prompt & Test Cases	Optimized Evaluation Matrix
AutoPrompt Shin et al. (2020)	Prompt Optimization	Gradientguided token search	Raw Prompt & Template	Optimized Prompt
APE Zhou et al. (2022)	Prompt Optimization	Twostage search loop	Task Description & Input-Output Example	Optimized Prompt
Context-Engineering Kim (2025)	Context Optimization	Heuristic & rulebased context orchestration	Raw Prompt & Context Components	Optimized Context (Prompt)
LlamaIndex Liu (2022)	Injection with RAG	Modular stack	Query & Knowledge Corpus	Context-augmented Output
LangChain RAG Chase (2022)	Injection with RAG	Chain graph wraps retriever	Query & Knowledge Corpus	Context-augmented Output
GraphRAG Larson et al. (2024)	Injection with RAG	LLMextracted entityrelation graph	Query & Knowledge Corpus	Context-augmented Output & Corpus KG
LightRAG Guo et al. (2024)	Injection with RAG	Stable incremental updates graph	Query & Knowledge Corpus	Context-augmented Output & Corpus KG
LlamaFactory Zheng et al. (2024)	Injection with PEFT	One-stop PEFT	Pretained LLMs & Instructions	Fine-tuned Model
Verl Team & Community (2024)	Injection with RL	One-stop PEFT	Pretrained LLMs & Instructions & RM	Fined-tuned Model
OpenRLHF Hu et al. (2024a)	Injection with RL	One-stop PEFT	Pretrained LLMs & Instructions	Fined-tuned Model
VLM-RL Huang et al. (2024d)	Injection with RL	One-stop RL	Pretrained VLMs & Instructions	Fine-tuned VLMs

3.2.1 EXPLICIT INJECTION

Given the centrality of prompts, explicit injection divides into two routes: knowledge-in-prompt and prompt optimization. Knowledge-in-prompt directly inserts domain contenttypically via RAG over unstructured sources (textbooks, papers, lab reports)to condition the model, with demonstrated systems in biomedicine (DALK Li et al. (2024a); BiomedRAG Li et al. (2024d)), finance (SNFinLLM Zhao et al. (2024a)), and chemistry (ChemCrow Bran et al. (2023); ChemAgent Tang et al. (2025b)). Prompt optimization, by contrast, enhances reasoning without adding new external content: Chain-of-Thought Wei et al. (2022) elicits intermediate steps, while automatic prompt-word optimization methodsPromptAgent Wang et al. (2024a) and Context-Engineering Kim (2025)tune prompts for specific downstream tasks.

3.2.2 Implicit Injection

To ensure the model's comprehension of domain-specific knowledge, implicit knowledge injection frequently leverages parameter updates, ensuring a more nuanced integration of domain expertise into the models functionality. Model adaptation for scientific agents spans supervised fine-tuning (SFT), reinforcement learning (RL), and modular knowledge adapters: SFT trains on inputoutput instruction pairs to yield well-defined outputs (e.g., scientific summaries, factual Q&A) while absorbing domain knowledge at scale; because it provides strong supervision, response quality is pivotal, and LLamaFactory supports efficient execution Xia et al. (2025); Li et al. (2025a); Horawalavithana et al. (2023); Zheng et al. (2024). For complex reasoning and preference alignment, RL employs RLVR with verifiable, outcome-based rewards tied to ground truth, RLHF with task-specific learned reward models, or DPO with implicit rewards from preferred vs. dispreferred pairs, with scalable implementations in Verl and OpenRLHF Narayanan et al. (2025a); Li et al. (2025a); Ren et al. (2025c); Team & Community (2024); Hu et al. (2024a). Complementarily, modular adapters inject domain knowledge while preserving base capabilitiesBioReason and ChatNT translate sequence attributes to support biological reasoning, and BLADE embeds specialized knowledge into smaller models that interface with large black-box models to boost downstream performance Fallahpour et al. (2025); de Almeida et al. (2025); Li et al. (2024b).

Table 4: Domain-specific tools examples of scientific agents.

Functionality	Tools	Calling Form	Domain	Target Descriptions	Typical Method
Expertise Acquisition	KEGG	REST API	Biology	Genome database	GeneGPT Jin et al. (2024b),
	UniProt/ PDB	REST API	Biology	Protein Sequence and Function Annotation Access	Biomni Huang et al. (2025a)
	NCBI	Python SDK	Biology& Medical	National Center for Biotechnology Information	GeneGPT Jin et al. (2024b), Crispr-GPT Huang et al. (2024a), BioResearcher Zhang et al. (2024b), et al.
	Semantic Scholar	REST API / Python SDK	General	Academic literature access	PaperQA Lála et al. (2023), KARMA Li & Wang (2025), et al.
	Arxiv	REST API	General	Scientific literature access	PaperQA Lála et al. (2023), Agent Laboratory Schmidgall et al. (2025), et al.
	Europe PMC	REST API	Medical	Scientific Literature & Metadata Access	LitChat Huang et al. (2025b)
Excution and Simulation	BLAST Camacho et al. (2009)	CLI / Python script	Biology	Highperformance Heuristicsbased Sequence Search	Crispr-GPT Huang et al. (2024a), GeneGPT Jin et al. (2024b)
	COPASI Hoops et al. (2006)	CLI / Python SDK	Biology	Biomedical Modeling	LLM-MC-Sim Zhao et al. (2023)
	AlphaFold Jumper et al. (2021)	Python script	Biology	Protein Structure Prediction	Virtual Lab Pak et al. (2024), AI co-scientist Gottweis et al. (2025)
	Vina / GNINA McNutt et al. (2021)	CLI / Python SDK	Chemistry	Molecular Docking	ChemCrow Bran et al. (2023), ChemAgent Tang et al. (2025b)
	Opentrons OT-2	Web GUI / Protocol API	Chemistry	Benchtop Liquid-handling Robot	Coscientist Boiko et al. (2023b)
	LabVIEW National Instruments (2024)	VI Graphical	General	Graphical Programming Environment	AILA Mandal et al. (2024)
	OpenFoam Weller et al. (2024)	CLI / C++ script	Physics	Fluid Dynamics & Heat Transfer Simulation	MetaOpenFoaM Chen et al. (2024a), Foam-Agent Yue et al. (2025)
	ROS	Python script	Robotics	Robotics Framework & Developer Toolkit	AILA Mandal et al. (2024)
Analysis and Visualization	Scanpy Wolf et al. (2018)	Python script	Biology	Large-scale Single-cell Gene Expression Data Analysis	CellAgent Xiao et al. (2024)
	DESeq2	R script	Biology	RNA-seq Data Analysis	IAN Nagarajan et al. (2025)
	PyMOL Sun et al. (2025a)	CLI / python script	Biology	Biomolecular Structure Visualization & Analysis	PyMOLfold Sun et al. (2025a)
	Seurat Hao et al. (2023)	R script	Biology	Single-cell Data Alignment	CellTypeAgent Chen et al. (2025a)
	Nextflow Tommaso et al. (2017)	DSL2 / CLI	Biology	Dataflow-based Workflow Engine	FlowAgent Shi et al. (2025)
	ChimeraX Meng et al. (2023)	CLI / Python script	Biology	Atomic Model Construction & Analysis	Virtual Lab Pak et al. (2024)
	RXNAiChem Schwaller et al. (2019)	Python script	Chemistry	Chemical Reaction Outcome Prediction	ChemCrow Bran et al. (2023)
	RDKit	Python script	Chemistry	Compound Analysis Using Machine Learning	Chemcrow Bran et al. (2023), Coscientist Boiko et al. (2023b)
	Ggplot2 Wickham (2016)	R script	General	Data Visualization	DrBioRight 2.0 Liu et al. (2025c)
Interaction	Streamlit	Web Socket / Python script	General	Web Application Construction	DrBioRight 2 Liu et al. (2025c)
	JupyterHub	REST API	General	Jupyter Environment Access	Jupybara Wang et al. (2025b)
	GitHub	REST API / Git Protocol	General	Collaborative Software Development	Agent Laboratory Schmidgall et al. (2025)
	Zenodo	REST API	General	Research Metadata Repository	AgentRxiv Schmidgall & Moor (2025), Agent Laboratory Schmidgall et al. (2025)

3.3 Tool Integration

Tool integration is fundamental to extending agents abilities along two axestool selection and integration implementation. For selection, Table 4 groups tools into four categories: (i) expertise acquisition, where domain knowledge bases integrate into RAG via LangChain and AutoGen Chase (2022); Wu et al. (2023); (ii) execution and simulation, encompassing SDKs/scripts and specialized models/automation Camacho et al. (2009); Macenski et al. (2022); Jumper et al. (2021); Boiko et al. (2023b); Opentrons Labworks Inc. (2024); (iii) analysis and visualization, spanning molecular graphics, plotting, and workflow orchestration Meng et al. (2023); Wickham (2016); Tommaso et al. (2017); and (iv) interaction, via collaborative interfaces and repositories Group (2025); Team (2025). For implementation, agent frameworks decompose tasks, plan tool use, and invoke tools in end-to-end pipelines that manifest as linear chains or tree-structured flows; most systems employ in-cycle planners such as CoT and ReAct Wei et al. (2022); Yao et al. (2023b), while complex problems favor branching plans with RL/MCTS-style search for backtracking and search-space optimizatione.g., AI Scientist-v2 and Robinwithin formalized tool-learning frameworks Qin et al. (2024); Qu et al. (2025); Lu et al. (2024); Ghareeb et al. (2025).

4 Scientific Agent Capabilities Enhancement

Following Section 3, to elevate the scientific agents to the level of avatar, targeted capability enhancement across multiple dimensions is essential. As illustrated in Fig. 3(b), we categorize these capabilities enhancement into three key dimensions: memory, reasoning, and collaboration.

4.1 Memory Enhancement

Memory is pivotal for human-level scientific agentssustaining context and continuity, enabling multi-step reasoning and planning, and supporting experiential learningyet most designs remain

context-bound; a few, e.g., ChemAgent Tang et al. (2025b), extend memory across planning, execution, and knowledge; accordingly, we outline enhancement strategies along two axesmemory structure and memory systemsummarized in Table 5.

4.1.1 Memory Sturucture

Different memory structures reflect different levels of memory knowledge. Currently, as discussed in Zeng et al. (2024a), the memory structure organization in LLM-based agents can be roughly divided into four categories: blocks, knowledge triples, atomic facts, and summaries.

Chunks represent contextual segments often cached or retrieved to simulate long-term memory, adopted by methods such as RAG Gao et al. (2024b), MemGPT Packer et al. (2024) and ThinkIn-Memory Liu et al. (2023a). This is the most widely used and original memory structure. Knowledge triples, inspired by Davidsonian semantic theory, model conceptual relations within sentences using a <event, subject, object> structure. This form of memory is adopted by systems like RET-LLM Modarressi et al. (2024) and AriGraph Anokhin et al. (2025), which encode event-based knowledge in a structured and interpretable format. Atomic facts encapsulate factual knowledge into minimal discrete memory units. This fine-grained decomposition facilitates fine-grained retrieval, as utilized in Graphreader Li et al. (2024e). Summaries condense large-scale context into human-like gist memory, as seen in Lee et al. (2024a). In addition, recent efforts have explored routine-based memory, where procedural patterns discovered during interaction are stored and reused. AFlow Zhang et al. (2024a), PGPO Cao et al. (2025b), and Agent Workflow Memory Wang et al. (2024d) capture and organize experience in the form of reusable action routines, allowing memory to generalize and transfer to other tasks.

Table 5: Scientific agents' memory enhancement methods.

Category	Method	Memory Structure	Implement	Task Settings
Context-Centirc	TRIME Zhong et al. (2022)	Chunks	Parameters update	Machine Translation and Language Modeling
	MemoryBank Zhong et al. (2023)	Summary	Memory repo	Long-term conversation
	IRCoT Trivedi et al. (2023)	Chunk	Memory repo	Multi-hop Q&A
	MemoChat Lu et al. (2023)	Atomic Facts	Parameters update & memorandum	STEM exams & literary writing
	MemGPT Packer et al. (2024)	Summary	Dynamic Memory Queue	Multi-session chat and document analysis
	AFlow Zhang et al. (2024a)	Routine	Operator & MCTS	General Tasks
Adaption-Centric	VOYAGER Wang et al. (2023a)	Skill Library	Curriculum design	Minecraft exploring
	FINMEM Yu et al. (2023)	Summary	Working Memory+Layered Long-Term Memory	Stock Trading
	HIAGENT Hu et al. (2024b)	Summary	Observation Summarization+Trajectory Retrieval	Long-term tasks in AgentBoard
	RET-LLM Modarressi et al. (2024)	Triples	Tool-use	Long-term conversation
	AriGraph Anokhin et al. (2025)	Triples	Memory Graph	Roguelike & Interactive games
	A-Mem Xu et al. (2025c)	Summary(Structured)	Note Construction+Link Generation+Memory Evolution	Long conversation Q&A and Multi-party Conversation
	G-Memory Zhang et al. (2025a)	Summary	Graph	Knowledge reasoning, embodied action, games

4.1.2 Memory Systems

Memory in scientific agents spans two archetypescontext-centric and action-centricwhere the former distills reliable facts from interaction histories to curb hallucinations and preserve context, while the latter treats memory as a dynamic process that abstracts skills from experience to improve planning and generalization to changing tasks. Context-centric systems include TRIME, which injects memory at training time to boost performance Zhong et al. (2022); MemoryBank, which organizes storage, retrieval, and updating via an Ebbinghaus-inspired schedule Zhong et al. (2023); ChatDB, a symbolic, relational store that strengthens multi-hop reasoning Hu et al. (2023); MemoChat, which maintains coherence through a MemoryRetrievalResponse loop Lu et al. (2023); and MemGPT, which adopts hierarchical, OS-like memory for long-range reasoning Packer et al. (2024). Operation-centric designs emphasize exploration and adaptation: VOYAGER continually learns reusable skills in an open world Wang et al. (2023a); FINMEM models human-like decisions with working/long-term financial memories Yu et al. (2023); HIAGENT uses subgoals as memory units with trajectory retrieval for long tasks Hu et al. (2024b); RET-LLM provides scalable, aggregatable, updateable, and interpretable memory units Modarressi et al. (2024); an LLM Agent Swarm for Drug Discovery shares memory across agents to avoid redundancy Song et al. (2025a); AriGraph fuses semantic and episodic memories in a hybrid graph Anokhin et al. (2025); A-Mem builds a Zettelkasten-inspired network for dynamic indexing/linking Xu et al. (2025c); and G-Memory constructs hierarchical, graph-based stores to manage complex histories in multi-agent systems Zhang et al. (2025a).

Table 6: Scientific agent collaboration challenge and mitigations.

Catagrization	Example Agents	Challenges	Mitigations
Rolespecialised agents	ChemAgents Polat et al. (2025), SciAgents Ghafarollahi & Buehler (2025a), AtomAgents Ghafarollahi & Buehler (2024), et al.	Role Overlap Information flow	Explicit role spe Hierarchical control
Dialog & debate	Virtual Lab Pak et al. (2024), AI Coscientist Gottweis et al. (2025), et al.	Discussion drift Conflict deadlocks	Structured protocols Voting agent to declare consensus
Knowledge sharing	SciAgents Ghafarollahi & Buehler (2025a), et al.	Stale or inconsistent shared facts data modalities	Versioned knowledge graph Unification layer for heterogeneous data
Goal setting & feedback	ChemAgents Polat et al. (2025),Sparks Ghafarollahi & Buehler (2025c), ChatGPT Research Group Zheng et al. (2023), et al.	Balancing oversight Ambiguous or delayed feedback loops	AI plan with human edits Inline annotation or comment system
Naturallanguage interface	ChemAgents Polat et al. (2025), MatPilot Ni et al. (2024), et al.	Linguistic ambiguity with implicit constraints Context drift in multiturn dialogue	Postprocessing & reformat responses Structured conversation memory

4.2 Reasoning Enhancement

Grounded, verifiable, and traceable reasoning underpins scientific-agent responses Li et al. (2025d), yet LLM hallucinations, inconsistencies, and contradictions erode reliability on complex tasks, motivating a progressive treatment of reasoning enhancementfrom general scientific reasoning to domain-specific optimization.

4.2.1 General Reasoning in Scientific Scenario

Reasoning capabilities transfer effectively to scientific research because reasoning operationalizes dense knowledge, enabling accurate decisions under overload and uncertainty; mainstream mechanisms center on structured reasoning chains and self-consistency verification (Fig. 5). Structured approaches span single-round Chain-of-Thought (CoT; Lets think step by step) to enhance reasoning via prompt engineering Wei et al. (2022), few-shot prompting to supply task-relevant exemplars Xu et al. (2023); Zhang et al. (2023b); Yasunaga et al. (2024), and multi-round decomposition that iteratively solves sub-problems Yao et al. (2023a); Zhou et al. (2023); Besta et al. (2024), with Buffer-of-Thoughts introducing reusable thought templates to balance generality and cost Yang et al. (2024a). Complementarily, self-consistency generates paraphrased prompts, samples diverse reasoning paths, and aggregates by majority vote to curb stochastic errors Wang et al. (2023d), further strengthened by multi-model critique and reconciliation Du et al. (2024b) and Reflexions structured linguistic feedback without full RL Shinn et al. (2023); together, these strategies reduce ambiguity and error propagation, improving robustness for scientific decision-making.

4.2.2 Domain-Specific Reasoning Optimization

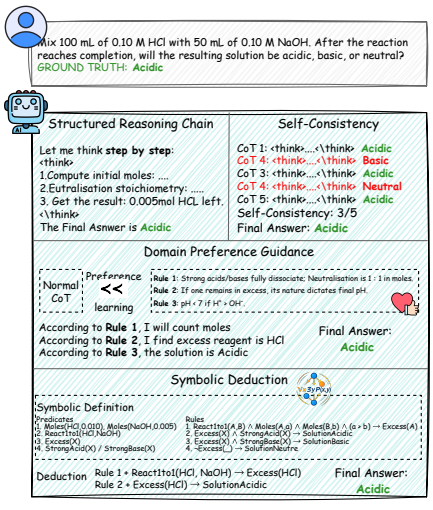


Figure 5: Illustration of scientific agent reasoning enhancement.

As agents progress to scientific inquiry, reasoning must bridge knowledge gaps and enforce strict scientific laws; a scientific augmentation layer addresses these via domain preference guidance and symbolic deduction to ensure reliable deduction. Domain preference guidance injects domain logic so models reason with discipline-specific priorsfor example, Ether0 cold-starts with annotated thinking chains before large-scale RL for compound-specification design Narayanan et al. (2025a); ChemCrow and ChatMOF integrate specialized external models to embed task knowledge into agent pipelines Bran et al. (2023); Kang & Kim (2024); and preference alignment/model editing further strengthens domain-aware reasoning Zhang et al. (2024c); Shahriar et al. (2024). Complementarily, symbolic deduction couples probabilistic LLMs with deterministic formalisms to create a closed-loop validator: code serves as an executable witness (e.g., CODEI/O converts reasoning to code Li et al. (2025b); PAL delegates solution steps to a program interpreter Gao et al. (2023a)); formal proof systems translate problems for verification (DeepSeek-Prover, Gödel-Prover, DeepSeek-Prover-v2 Xin et al. (2024); Lin et al. (2025); Ren et al. (2025c)); and domain formalisms extend beyond math (formalized adjuvant logic Yi et al. (2025) and inductive reflection for neural-symbolic reasoning in ABL-Refl Hu et al. (2025)).

4.3 Collaboration Enhancement

Scientific research is evolving with advanced frameworks integrating multi-agent systems and human collaboration Song et al. (2025b); Zheng et al. (2023); Ghafarollahi & Buehler (2025b); Schmidgall & Moor (2025). This marks a shift from old monolithic tools to dynamic "agent laboratories", where AI agents and humans collaborate. Orchestrated by LLMs, these systems have modular architectures for interaction and knowledge exchange, succeeding in domains like chemistry Song et al. (2025b); Zheng et al. (2023), materials science Yin et al. (2025); Ni et al. (2024), and biology Gao et al. (2024a). The sophisticated design of collaborative paradigms has enhanced the capabilities of intelligent agents in complex tasks and promoted their implementation. Table 6 further refines these categories for the corresponding collaborative enhancement paradigms and lists the challenges faced and corresponding mitigation strategies.

4.3.1 Multi-Agents Collaboration

In multi-agent scientific systems, collaboration advantages cluster into three dimensionsrole specialization, dialogue/debate, and knowledge sharing. Explicit, hierarchical role design reduces per-agent complexity and boosts throughput (e.g., ChemAgents for collaborative experiment planning/execution; seven-assistant lab roles in the ChatGPT Research Group; domain roles in SciAgents and AtomAgents) while LLMs help interpret tasks and adapt roles to shifting needs, though pitfalls remain around clear boundaries and smooth information flow Song et al. (2025b); Zheng et al. (2023); Ghafarollahi & Buehler (2025a; 2024). Effective communication relies on structured meetings, agenda setting, and critique cycles to keep debates coherent and convergent, as in Virtual Labs PI-guided teams and the AI co-scientists generatedebateevolve loop for hypothesis refinement Kudiabor (2024); Gottweis et al. (2025). Finally, shared knowledgevia centralized graphs and common workspacesunderpins mutual understanding, consistent updates, and distributed reasoning; SciAgents illustrates how shared repositories can unlock discoveries beyond a single agents capacity Ghafarollahi & Buehler (2025a).

4.3.2 Human-AI Collaboration

Human expertise remains indispensable in mixed-initiative scientific systems, where humans and AI agents collaborate synergistically by combining strategic guidance with scalable automation. Effective designs translate high-level human goals into tractable agent tasks and incorporate imprecise feedback without drifting into micromanagement; concrete implementations include ChemAgents Task Manager for human objective setting and progress review, the ChatGPT Research Groups dialogue-driven goal setting, and Agent Laboratorys feedback-rich, alignment-aware workflows Song et al. (2025b); Zheng et al. (2023); Schmidgall et al. (2025). Equally critical are natural-language interfaces powered by LLMs/LMMs that reduce access barriers while handling ambiguity, intent parsing, and clear response generatione.g., ChemAgents on-premises LLM integration and MatPilots conversational hypothesis/experiment designthereby tightening the humanAI loop and broadening participation across the scientific community Song et al. (2025b); Ni et al. (2024).

5 Conclusion

This article presents the Hitchhikers Guide to Autonomous Research – a comprehensive survey of scientific agents. Initially, we provide an overview of these agents, contrasting them with general agents, and taxonomizing them into three advancing levels: agent as assistant, agent as partner, and agent as avatar. Subsequently, we offer a details guide to the construction process of scientific agents, elucidating how to construct agents from foundational principles through knowledge organization, knowledge injection, and tool integration. The discussion extends to enhancing agents concerning memory, reasoning, and collaboration. Additionally, the paper reviews existing benchmarks and mainstream evaluation indicators pertinent to scientific agents. Ultimately, it anticipates future research trajectories within the domain of scientific agents. This review aspires to stimulate scientists across various fields to develop scientific agents, thereby fostering the enduring advancement of AI-driven autonomous research. Finally, we maintain a real-time repository Awesome_Scientific_Agent to monitor the latest advancements in scientific agents.

REFERENCES

Abbi Abdel-Rehim, Hector Zenil, and et al. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of the Royal Society Interface*, 22(227):20240674, 2025.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

Samuel Alber, Bowen Chen, Eric Sun, Alina Isakova, Aaron James Wilk, and James Zou. Cellvoyager: Ai compbio agent generates new insights by autonomously analyzing biological data. *bioRxiv*, pp. 2025–06, 2025.

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. Arigraph: Learning knowledge graph world models with episodic memory for llm agents. *arXiv preprint arXiv:2407.04363*, 2025.

Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku, 2024. Technical report.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Xuefeng Bai, Song He, Yi Li, Yabo Xie, Xin Zhang, Wenli Du, and Jian-Rong Li. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials*, 11(1):51, 2025.

Adib Bazgir, Yuwen Zhang, et al. Multicrossmodal automated agent for integrating diverse materials science data. *arXiv preprint arXiv:2505.15132*, 2025.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *EMNLP*, 2019. doi: 10.48550/arXiv.1903.10676.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI 2024*, volume 38, pp. 17682–17690, 2024.

Christopher M. Bishop. AI4Science to Empower the Fifth Paradigm of Scientific Discovery, 2022. Blog post, July 7.

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023a.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023b.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Christiam Camacho, George Coulouris, Valery Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009. doi: 10.1186/1471-2105-10-421.

Christian Cao, Rohit Arora, Paul Cento, Katherine Manta, Elina Farahani, Matthew Cecere, Anabel Selemon, Jason Sang, Ling Xi Gong, Robert Kloosterman, et al. Automation of systematic reviews with large language models. *medRxiv*, 2025a. doi: 10.1101/2025.06.13.25329541.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.

Zouying Cao, Runze Wang, Yifei Yang, Xinbei Ma, Xiaoyong Zhu, Bo Zheng, and Hai Zhao. Pgpo: Enhancing agent reasoning via pseudocodestyle planning guided preference optimization. *arXiv preprint arXiv:2506.01475*, 2025b.

Franck Cappello, Sandeep Madireddy, Robert Underwood, et al. Eaira: Establishing a methodology for evaluating ai models as scientific research assistants. *arXiv preprint arXiv:2502.20309*, 2025.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mdry. Mlebench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2025.

Harrison Chase. Langchain: Building applications with llms through composability, 2022. Accessed: 2025-07-12.

Jiawen Chen, Jianghao Zhang, Huaxiu Yao, and Yun Li. Celltypeagent: Trustworthy cell type annotation with large language models. *arXiv*, abs/2505.08844, May 2025a. Preprint.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, et al. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*, 2025b.

Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. Autobench: An automated benchmark for scientific discovery in llms. *arXiv preprint arXiv:2502.15224*, 2025c.

Yuxuan Chen, Xu Zhu, Hua Zhou, and Zhuyin Ren. Metaopenfoam: an llmbased multiagent framework for cfd. *arXiv*, abs/2407.21320, Jul 2024a. Preprint.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024b.

Ashwin Chithra, Brian Goodall, Phillip Pope, et al. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

promptfoo community. PromptFoo: Test your prompts, agents, and rags a developerfriendly llm evaluation toolkit, 2025. Opensource software; see GitHub for documentation.

Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, et al. Organa: a robotic assistant for automated chemistry experimentation and characterization. *Matter*, 8(2), 2025.

Bernardo P de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, et al. A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*, pp. 1–14, 2025.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, and *et al.* DeepSeek-V3 Technical Report, 2024.

Jiangshu Du, Yibo Wang, and et al. Llms assist nlp researchers: Critique paper (meta-) reviewing. *arXiv preprint arXiv:2406.16253*, 2024a.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025a.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025b.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML 2024*, 2024b.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J Maddison, et al. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *arXiv preprint arXiv:2505.23579*, 2025.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. Citebench: A benchmark for scientific citation text generation. *arXiv preprint arXiv:2212.09577*, 2022.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *ICML 2023*, pp. 10764–10799. PMLR, 2023a.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *EMNLP 2023*, 2023b.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024b.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*, 2024c.

Gemini Team, Google DeepMind. Gemini: A Family of Highly Capable Multimodal Models, 2023. Technical report.

Alireza Ghafarollahi and Markus J. Buehler. Atomagents: Alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. *arXiv preprint arXiv:2407.10022*, 2024.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025a.

Alireza Ghafarollahi and Markus J Buehler. Automating alloy design and discovery with physics-aware multimodal multiagent ai. *Proceedings of the National Academy of Sciences*, 122(4): e2414074122, 2025b.

Alireza Ghafarollahi and Markus J. Buehler. Sparks: Multi-agent artificial intelligence model discovers protein design principles. *arXiv preprint arXiv:2504.19017*, 2025c.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Jon M Laurent, Muhammed T Razzak, Andrew D White, Michaela M Hinks, and Samuel G Rodriques. Robin: A multi-agent system for automating scientific discovery. *arXiv preprint arXiv:2505.13400*, 2025.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Mourad Gridach, Jay Nanavati, Christina Mack, Khaldoun Zine El Abidine, and Lenon Mendes. Agentic AI for scientific discovery: A survey of progress, challenges, and future directions. In *ICLR 2025 Workshop AgenticAI*, 2025.

CERN Open Data Group. Zenodo: General-purpose open-access repository. online platform, 2025. Assigns DOIs to research outputs.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2025.

Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Anah-v2: Scaling analytical hallucination annotation of large language models. In *NeurIPS 2024*, 2024.

Haiyang Guo, Fanhu Zeng, Fei Zhu, Jiayi Wang, Xukai Wang, Jingang Zhou, Hongbo Zhao, Wenzhuo Liu, Shijie Ma, Xu-Yao Zhang, et al. A comprehensive survey on continual learning in generative models. *arXiv preprint arXiv:2506.13045*, 2025.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. Medalpaca: An open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Yuhan Hao, Tim Stuart, Madeline H. Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023. doi: 10.1038/s41587-023-01767-y.

Dan Hendrycks, Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, and et al. Humanitys last exam: A hard benchmark at the frontier of human knowledge. *arXiv preprint arXiv:2501.14249*, 2025.

S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. Copasi: a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006. doi: 10.1093/bioinformatics/btl485.

Sameera Yasanka Horawalavithana, Sai Munikoti, Ian B. Stewart, and Henry J. Kvinge. Scitune: Aligning large language models with scientific multimodal instructions. *arXiv preprint arXiv:2307.01139*, 2023.

Ting-Yao Hsu, C. Lee Giles, and Ting-Hao 'Kenneth' Huang. Scicap: Generating captions for scientific figures. In *EMNLP 2021 Findings*, pp. 3258–3264. Association for Computational Linguistics, November 2021.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework, 2024a.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. *arXiv preprint arXiv:2408.09559*, 2024b.

Wen-Chao Hu, Wang-Zhou Dai, Yuan Jiang, and Zhi-Hua Zhou. Efficient rectification of neuro-symbolic reasoning inconsistencies by abductive reflection. In *AAAI 2025*, volume 39, pp. 17333–17341, 2025.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024a.

Kexin Huang, Serena Zhang, Hanchen Wang, and et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, Jun 2025a. doi: 10.1101/2025.05.30.656746. Preprint.

Mingyu Huang, Shasha Zhou, Yuxuan Chen, and Ke Li. Conversational exploration of literature landscape with litchat. *arXiv*, abs/2505.23789, May 2025b. Preprint.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2024b.

Zhiheng Huang, Hao Zhou, Juncheng Li, and et al. Drugagent: Automating ai-aided drug discovery with multi-agent collaboration. *arXiv preprint arXiv:2411.15692*, 2024c. doi: 10.48550/arXiv.2411.15692.

Zilin Huang, Zihao Sheng, Yansong Qu, Junwei You, and Sikai Chen. Vlm-rl: A unified vision language models and reinforcement learning framework for safe autonomous driving. *arXiv preprint arXiv:2412.15544*, 2024d.

Kevin M. Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6:161–169, 2024. doi: 10.1038/s42256-023-00788-1.

Peter Jansen, MarcAlexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Ø yvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *arXiv preprint arXiv:2406.06769*, 2024.

Shuyi Jia, Chao Zhang, and Victor Fung. LLMatDesign: Autonomous materials discovery with large language models. *arXiv preprint arXiv:2406.13163*, 2024. doi: 10.48550/arXiv.2406.13163.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Di Jin, Eileen Pan, Nassim Oufattole, WeiHung Weng, and Hanyi Fang. What disease does this patient have? a largescale open domain question answering dataset from medical exams. *Applied Sciences*, 11(2):732, 2021. doi: 10.3390/app11020732.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W. J. Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and Zhiyong Lu. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *arXiv preprint arXiv:2402.13225*, 2024a. doi: 10.48550/arXiv.2402.13225.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 2024b. doi: 10.1093/bioinformatics/btae075.

J. Jumper, R. Evans, A. Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

Yeonghun Kang and Jihan Kim. Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature communications*, 15(1): 4705, 2024.

David Kim. Context engineering: A practical handbook for context design, orchestration, and optimization, 2025.

Patrick Tser Jern Kon, Jiachen Liu, et al. Exp-bench: Can ai conduct ai research experiments? *arXiv preprint arXiv:2505.24785*, 2025.

Helena Kudiabor. Virtual lab powered by'ai scientists' super-charges biomedical research. *Nature*, 636(8043):532–533, 2024.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. Sciclaimhunt: A large dataset for evidencebased scientific claim verification. *arXiv preprint arXiv:2502.10003*, 2025.

Varun V. Kumar, Leonard Gleyzer, Adar Kahana, Khemraj Shukla, and George E. Karniadakis. Mycrunchgpt: A chatgpt assisted framework for scientific machine learning. *arXiv preprint arXiv:2306.15551*, 2023.

LangChain Inc. Langgraph: A stateful orchestration framework for multi-actor agent applications, 2024. MIT License; Accessed: 20250712.

Jonathan Larson, Steven Truitt, Darren Edge, Ha Trinh, Bryan Li, Alonso Guevara Fernández, and Joshua Bradley. Graphrag: A modular graph-based retrieval-augmented generation (rag) system, 2024.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. Labbench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*, 2024a.

Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Transactions on Intelligent Systems and Technology*, 2024b.

Chen Li et al. Chemau: Harnessing the reasoning of large language models in chemical research via instruction fine-tuning on qwen-1.5 b. *arXiv preprint arXiv:2502.01234*, 2025a.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, Huan Liu, Li Shen, and Tianlong Chen. DALK: Dynamic co-augmentation of LLMs and KG to answer Alzheimer's disease questions with scientific literature. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *EMNLP 2024 Findings*, pp. 2187–2205, November 2024a.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Zhijing Wu, Yiqun Liu, Chong Chen, and Qi Tian. Blade: Enhancing black-box large language models with small domain-specific models, 2024b.

Junlong Li, Daya Guo, Dejian Yang, Runxin Xu, Yu Wu, and Junxian He. Codei/o: Condensing reasoning patterns via code input-output prediction. *arXiv preprint arXiv:2502.07316*, 2025b.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large visionlanguage models. In *Proceedings of the 62nd Annual Meeting of the ACL (Long Papers)*, pp. 14369–14387, 2024c. doi: 10.18653/v1/2024.acl-long.775.

M Li, H Kilicoglu, H Xu, and R Zhang. Biomedrag: A retrieval augmented large language model for biomedicine. arxiv. *arXiv preprint arXiv:2405.00465*, 2024d.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. *arXiv preprint arXiv:2406.14550*, 2024e.

Xin Li, Chen Chen, Ruocheng Guo, and et al. Iris: Intelligent research interaction suite for multimodal scientific workflows. *arXiv preprint arXiv:2502.05811*, 2025c.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.

Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. Chatcite: Llm agent with human workflow guidance for comparative literature summary. *arXiv preprint arXiv:2403.02574*, 2024f.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, ... Stephen D. Wilson, Woosang Lim, and William Yang Wang. Mmsci: A multimodal multidiscipline dataset for phdlevel scientific comprehension. *arXiv preprint arXiv:2407.04903*, 2024g. doi: 10.48550/arXiv.2407.04903.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025d.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z. Li, and Kaicheng Yu. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science, 2024.

Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia Li, Mengzhou Xia, Danqi Chen, Sanjeev Arora, et al. Goedel-prover: A frontier model for open-source automated theorem proving. *arXiv preprint arXiv:2502.07640*, 2025.

Chengwei Liu, Chong Wang, Jiayue Cao, Jingquan Ge, Kun Wang, Lyuye Zhang, Ming-Ming Cheng, Penghai Zhao, Tianlin Li, Xiaojun Jia, et al. A vision for auto research with llm agents. *arXiv preprint arXiv:2504.18765*, 2025a.

Haoyang Liu and Haohan Wang. Genotex: A benchmark for evaluating llm-based exploration of gene expression data in alignment with bioinformaticians. *arXiv preprint arXiv:2406.15341*, 2024.

Haoyang Liu, Yijiang Li, Jinglin Jian, Yuxuan Cheng, Jianrong Lu, Shuyi Guo, Jinglei Zhu, Mianchen Zhang, Miantong Zhang, and Haohan Wang. Toward a team of ai-made scientists for scientific discovery from gene expression data. *arXiv preprint arXiv:2402.12391*, 2024a.

Jerry Liu. Llamaindex: A data framework for llm applications, 2022.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*, 2023a.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*, 2022.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023b.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration, 2025b.

Wei Liu, Jun Li, Yitao Tang, Yining Zhao, Chaozhong Liu, Meiyi Song, Zhenlin Ju, Shwetha V Kumar, Yiling Lu, Rehan Akbani, et al. Drbioright 2.0: an llm-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis. *Nature communications*, 16(1):2256, 2025c.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023c.

Xinyi Liu, Lipeng Ma, Yixuan Li, Weidong Yang, Qingyuan Zhou, Jiayi Song, Shuhao Li, and Ben Fei. Chemau: Harness the reasoning of llms in chemical research with adaptive uncertainty estimation. *arXiv preprint arXiv:2506.01116*, 2025d.

Zheng Liu, Wenhao Li, Qiuqiang Kong, and et al. Protagents: Protein discovery via large language model multi-agent collaboration. *arXiv preprint arXiv:2402.04268*, 2024b. doi: 10.48550/arXiv.2402.04268.

Chris Lu, Cong Lu, Robert T. Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. doi: 10.48550/arXiv.2408.06292.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*, 2023.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, KaiWei Chang, SongChun Zhu, Øyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS 2022*, 2022.

Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, James Zou, and OctoTools Contributors. Octotools: An agentic framework with extensible tools for complex reasoning. *Preprint on arXiv and open-source code on GitHub*, 2025.

Yuxing Lu and Jinzhuo Wang. Karma: Leveraging multi-agent llms for automated knowledge graph enrichment. *arXiv*, abs/2502.06472, Feb 2025. Preprint.

Junyu Luo, Weizhi Zhang, Ye Yuan, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 24(1):bbac409, 2023.

Yi Luo, Linghang Shi, Yihao Li, Aobo Zhuang, Yeyun Gong, Ling Liu, and Chen Lin. From intention to implementation: Automating biomedical research via llms. *arXiv preprint arXiv:2412.09429*, 2024.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall. Robot operating system 2: Design, architecture, and uses in the wild. *Science Robotics*, 7(66):eabm6074, 2022. doi: 10.1126/scirobotics.abm6074.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, 20(4):647–665, 2016.

Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M. Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and N. M. Anoop Krishnan. Autonomous microscopy experiments through large language model agents. *arXiv*, abs/2501.10385, Dec 2024. Preprint.

A. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, and D. R. Koes. Gnina 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 2021. doi: 10.1186/s13321-021-00522-2.

Elaine C. Meng, Thomas D. Goddard, Eric F. Pettersen, Gregory S. Couch, Zach J. Pearson, Jonathan H. Morris, and Thomas E. Ferrin. Ucsf chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023. doi: 10.1002/pro.4792.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-llm: Towards a general read-write memory for large language models. *arXiv preprint arXiv:2305.14322*, 2024.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *NeurIPS 2021*, 2021.

João Moura and CrewAI Contributors. Crewai: A lean, lightningfast python framework for multiagent orchestration, 2024. Accessed: 20250712; MIT License.

Vijayaraj Nagarajan, Guangpu Shi, and et al. Ian: An intelligent system for omics data analysis and discovery. *bioRxiv*, 2025.

Siddharth Narayanan, James Braza, Albert Bou, Geemi Wellawatte, Mayk Caldas, Ludovico Mitchener, Samuel G. Rodriques, and Andrew D. White. ether0: A scientific reasoning model for chemistry. Website, 2025a.

Siddharth Narayanan, James D. Braza, Ryan-Rhys Griffiths, Manu Ponnapati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Aviary: Training language agents on challenging scientific tasks. *arXiv preprint arXiv:2412.21154*, 2025b. doi: 10.48550/arXiv.2412.21154.

National Instruments. *LabVIEW: Laboratory Virtual Instrument Engineering Workbench*. National Instruments, 2024. Latest stable version 2024Q3 (v24.3), originally released 1986; proprietary.

Tuan Dung Nguyen, Yuan-Sen Ting, et al. Astrollama-chat: Scaling astrollama with conversational and diverse datasets. *arXiv preprint arXiv:2401.12345*, 2024.

Ziqi Ni, Yahao Li, Kaijia Hu, Kunyuan Han, Ming Xu, Xingyu Chen, Fengqi Liu, Yicong Ye, and Shuxin Bai. Matpilot: an llm-enabled ai materials scientist under the framework of human-machine collaboration. *arXiv preprint arXiv:2411.08063*, 2024.

Zhangming Niu, Xianglu Xiao, and et al. Pharmabench: Enhancing admet benchmarks with large language models. *Scientific Data*, 11(985), 2024. doi: 10.1038/s41597-024-03793-0.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Charles O'Neill, Tirthankar Ghosal, Roberta Rileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciuc. Sparks of science: Hypothesis generation using structured paper data. *arXiv preprint arXiv:2504.12976*, 2025. doi: 10.48550/arXiv.2504.12976.

OpenAI. ChatGPT: Optimizing Language Models for Dialogue, 2022. Blog post, November 30.

Opentrons Labworks Inc. Ot2: Accessible benchtop liquidhandling robot. Product website, 2024. Released 2018; widely used in academic and COVID19 testing pipelines.

Shuyin Ouyang, Dong Huang, Jingwen Guo, Zeyu Sun, Qihao Zhu, and Jie M. Zhang. Dsbench: A realistic benchmark for data science code generation. *arXiv preprint arXiv:2505.15621*, 2025.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2024.

J. E. Pak, K. Swanson, W. Wu, N. L. Bulaong, et al. The virtual lab: Ai agents design new sarscov2 nanobodies with experimental validation. *bioRxiv*, 2024.11.11.623004, Nov 2024. doi: 10.1101/2024.11.11.623004. Preprint.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, pp. 311–318, 2002.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. BiMediX: Bilingual medical mixture of experts LLM. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16984–17002, November 2024.

Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. Position: Episodic memory is the missing piece for long-term llm agents, 2025.

Can Polat, Mehmet Tuncel, Hasan Kurban, Erchin Serpedin, and Mustafa Kurban. xchemagents: Agentic ai for explainable quantum chemistry. *arXiv preprint arXiv:2505.20574*, 2025. doi: 10.48550/arXiv.2505.20574.

Yingming Pu, Tao Lin, and Hongyu Chen. Piflow: Principle-aware scientific discovery with multi-agent collaboration. *arXiv preprint arXiv:2505.15047*, 2025.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. Tool learning with foundation models. *ACM Computing Surveys*, 57(4):1–40, 2024.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level googleproof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025a.

Z. Z. Ren, Zhihong Shao, Junxiao Song, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025b.

ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025c.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. Scicode: A scientist-curated benchmark for scientific code generation. *arXiv preprint arXiv:2407.13168*, 2024.

Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*, 2025.

Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

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

Philippe Schwaller, Teodoro Thakkar, Ola Engkvist, and Jean-Louis Reymond. Ibm rxn for chemistry: A chemical reaction prediction platform. online platform, 2019. IBM Molecular Transformer; Chem. Sci., 10, 370377.

Sadat Shahriar, Zheng Qi, Nikolaos Pappas, Srikanth Doss, Monica Sunkara, Kishaloy Halder, Manuel Mager, and Yassine Benajiba. Inference time llm alignment in single and multidomain preference spectrum. *arXiv preprint arXiv:2410.19206*, 2024.

Mahsa Sheikholeslami, Navid Mazrouei, Yousof Gheisari, Afshin Fasihi, Matin Irajpour, and Ali Motahharynia. Druggen enhances drug discovery with large language models and reinforcement learning. *Scientific Reports*, 15(1):13445, 2025. doi: 10.1038/s41598-025-98629-1.

Yuchen Shi, Siqi Cai, Zihan Xu, Yuei Qin, Gang Li, Hang Shao, Jiawei Chen, Deqing Yang, Ke Li, and Xing Sun. Flowagent: Achieving compliance and flexibility for workflow agents. *arXiv preprint arXiv:2502.14345*, 2025.

Hyungyu Shin, Jingyu Tang, et al. Mind the blind spots: A focus-level evaluation framework for llm reviews. *arXiv preprint arXiv:2502.17086*, 2025.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP 2020*, 2020.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS 2023*, volume 36, pp. 8634–8652, 2023.

Yoav Shoham. Agent-oriented programming. *Artificial intelligence*, 60(1):51–92, 1993.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *ICLR 2025*, 2025a. Oral Presentation, top 1.8%.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025b.

Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. Corebench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024.

Kevin Song, Andrew Trotter, and Jake Y. Chen. Llm agent swarm for hypothesis-driven drug discovery. *arXiv preprint arXiv:2504.17967*, 2025a.

Tao Song, Man Luo, Xiaolong Zhang, Linjiang Chen, Yan Huang, Jiaqi Cao, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, et al. A multiagent-driven robotic ai chemist enabling autonomous chemical research on demand. *Journal of the American Chemical Society*, 147(15):12534–12545, 2025b.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Chan Jun Shern, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, Tejal Patwardhan, and et al. Paperbench: Evaluating ais ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Jinyuan Sun et al. Interactive protein and ligand structure prediction in pymol. *arXiv*, abs/2502.00508, Feb 2025a. Preprint.

Liangtai Sun, Yang Han, and et al. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *AAAI 2024*, volume 38, pp. 19053–19061, 2024.

Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, Jianing Wang, Qintong Li, Xiangru Tang, Tianbao Xie, Xiachong Feng, Xiang Li, Ben Kao, Wenhai Wang, Biqing Qi, Lingpeng Kong, and Zhiyong Wu. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. *arXiv preprint arXiv:2505.19897*, 2025b.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025a.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2024. doi: 10.48550/arXiv.2311.10537.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, et al. Chemagent: Self-updating memories in large language models improves chemical reasoning. In *ICLR*, 2025b.

ByteDance Seed Team and Verl Community. Verl: Volcano engine reinforcement learning for large language models, 2024. Accessed: 2025-07-14.

Jupyter Development Team. Jupyterhub. software platform, 2025. Multi-user Jupyter notebook server.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*, 2024.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, et al. Ai research agents for machine learning: Search, exploration, and generalization in mle-bench. *arXiv preprint arXiv:2507.02554*, 2025.

Paolo Di Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35:316–319, 2017. doi: 10.1038/nbt.3820.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *ACL 2023*, pp. 10014–10037, July 2023.

Cunshi Wang, Xinjie Hu, Yu Zhang, and et al. Starwhisper telescope: Agent-based observation assistant system to approach an ai astrophysicist. *arXiv preprint arXiv:2412.06412*, 2025a. doi: 10.48550/arXiv.2412.06412.

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

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023b.

Huichen Will Wang, Larry Birnbaum, and Vidya Setlur. Jupybara: Operationalizing a design space for actionable data analysis and storytelling with llms. In *CHI 2025*, pp. 1005. ACM, April 2025b. doi: 10.1145/3706598.3713913.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259*, 2023c.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. Promptagent: Strategic planning with language models enables expert-level prompt optimization. In *ICLR*, 2024a. doi: 10.48550/arXiv.2310.16427. arXiv:2310.16427.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR 2023*, 2023d.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS 2024*, 2024b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS 2024*, 2024c.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS 2022*, 35:24824–24837, 2022.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*, 2024.

Henry Weller, Hrvoje Jasak, and Chris Greenshields. *OpenFOAM: The Open Source CFD Toolbox*, 2024. Version 12, The OpenFOAM Foundation, GPL3.0 licensed; latest release July 9, 2024.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *EMNLP 2024*, 2024. EMNLP 2024 Industry Track paper; originally arXiv:2411.00816.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: largescale singlecell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

Linjun Wu, Yujia Shen, Shang-Wen Zha, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2309.11474*, 2023.

Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, et al. Naturelm: Deciphering the language of nature for scientific discovery. *arXiv preprint arXiv:2502.07527*, 2025.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, and Jiajie Peng. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *arXiv*, abs/2407.09811, Jul 2024. Preprint.

Tong Xie, Yuwei Wan, Yixuan Liu, et al. DARWIN 1.5: Large language models as materials-science foundation models. *arXiv preprint arXiv:2412.11970*, 2024.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

Baixuan Xu, Chunyang Li, Weiqi Wang, Wei Fan, Tianshi Zheng, Haochen Shi, Tao Fan, Yangqiu Song, and Qiang Yang. Towards multi-agent reasoning systems for collaborative expertise delegation: An exploratory design study. *arXiv preprint arXiv:2505.07313*, 2025a.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*, 2023.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

Wei Xu, Gang Luo, Weiyu Meng, Xiaobing Zhai, Keli Zheng, Ji Wu, Yanrong Li, Abao Xing, Junrong Li, Zhifan Li, et al. Mragent: an llm-based automated agent for causal knowledge discovery in disease via mendelian randomization. *Briefings in Bioinformatics*, 26(2):bbaf140, 2025b.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025c.

Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4), 2021.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629*, 2025.

Fei Yang et al. Biomedgpt: A generalist vision–language foundation model for biomedical tasks. *arXiv preprint arXiv:2306.12345*, 2023.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. In *NeurIPS 2024*, volume 37, pp. 113519–113544, 2024a.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *arXiv preprint arXiv:2410.07076*, 2024b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *NeurIPS 2023*, 36:11809–11822, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR 2023*, 2023b.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. Large language models as analogical reasoners. In *ICLR 2024*, 2024.

Chen Yi, Zhang Yu, Xu Jian, Zhang XuYao, Yue Hua, Wang Xinming, Lyu Zequan, Wei Wei, and Liu ChengLin. Advancing adjuvant research with mllms: An open-ended benchmark and formal framework. *arXiv preprint*, 2025.

Ruoyan Avery Yin, Zhichu Ren, Zongyou Yin, Zhen Zhang, So Yeon Kim, Chia-Wei Hsu, and Ju Li. Collaborative ai enhances image understanding in materials science. *arXiv preprint arXiv:2503.13169*, 2025.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*, 2024.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *arXiv preprint arXiv:2311.13743*, 2023.

Ling Yue, Nithin Somasekharan, Yadi Cao, and Shaowu Pan. Foamagent: Towards automated intelligent cfd workflows. *arXiv*, abs/2505.04997, May 2025. Preprint.

Ruihong Zeng, Jinyuan Fang, Siwei Liu, and Zaiqiao Meng. On the structural memory of llm agents. *arXiv preprint arXiv:2412.15266*, 2024a.

Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. Mr-ben: A comprehensive meta-reasoning benchmark for large language models. *arXiv preprint arXiv:2406.13975*, 2024b.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*, 2025a.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. In *ICLR 2025*, 2025b.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*, 2024a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2020.

Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and Qiaozhu Mei. Massw: A new dataset and benchmark tasks for aiassisted scientific workflows. In *NAACL 2025 Findings*, pp. 2373–2394, 2025c. doi: 10.18653/v1/2025.findings-naacl.127.

Yan Zhang, Patrick Lewis, Qing Lyu, and et al. Automating biomedical research with a domain-specialized agent family. *arXiv preprint arXiv:2405.04422*, 2024b.

Yichi Zhang, Zhuo Chen, Yin Fang, Yanxi Lu, Li Fangming, Wen Zhang, and Huajun Chen. Knowledgeable preference alignment for LLMs in domain-specific question answering. In *ACL 2024 Findings*, pp. 891–904, August 2024c.

Yifan Zhang, Cheng Wei, Shangyou Wu, Zhengting He, and Wenhao Yu. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*, 2023a. doi: 10.48550/arXiv.2307.07930.

Yu Zhang, Yang Han, Shuai Chen, Ruijie Yu, Xin Zhao, Xianbin Liu, Kaipeng Zeng, Mengdi Yu, Jidong Tian, Feng Zhu, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. Large language models to accelerate organic chemistry synthesis. *arXiv*, abs/2504.18340, Apr 2025d. doi: 10.48550/arXiv.2504.18340. Preprint.

Zhongyue Zhang, Zijie Qiu, Yingcheng Wu, Sitan Li, Dingyan Wang, Zhuomin Zhou, Duo An, Yuhan Chen, Haijun Yu, Yongbo Wang, C.-Y. Ou, Zichen Wang, J. Chen, Bo Zhang, Yiwen Hu, Wenxin Zhang, ZhiJian Wei, Runze Ma, Qingwu Liu, Bo Dong, et al. Origene: A selfevolving virtual disease biologist automating therapeutic target discovery. *bioRxiv*, Jun 2025e. doi: 10.1101/2025.06.03.657658. Preprint.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR 2023*, 2023b.

Haiteng Zhao, Chang Ma, Fangzhi Xu, Lingpeng Kong, and Zhi-Hong Deng. Biomaze: Benchmarking and enhancing large language models for biological pathway reasoning. *arXiv preprint arXiv:2502.16660*, 2025.

Kai Zhao, Sheng Di, Xin Liang, Sihuan Li, Dingwen Tao, Julie Bessac, Zizhong Chen, and Franck Cappello. Sdrbench: Scientific data reduction benchmark for lossy compressors. *Proceedings of the Workshop on Emerging HPC Applications Supported by RIKEN, AIST, and Tokyo Tech*, 2021.

Shujuan Zhao, Lingfeng Qiao, Kangyang Luo, Qian-Wen Zhang, Junru Lu, and Di Yin. Snfinllm: Systematic and nuanced financial domain adaptation of chinese large language models. *arXiv preprint arXiv:2408.02302*, 2024a.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, et al. Chemdfm: A large language foundation model for chemistry. *arXiv preprint arXiv:2401.14818*, 2024b.

Zirui Zhao, Wee Sun Lee, and David Hsu. Llm-mc-sim: Enhancing large language models with monte carlo simulation for complex reasoning tasks. *arXiv*, abs/2305.14078, May 2023. Preprint.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259*, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llama factory: Unified efficient fine-tuning of 100+ language models, 2024.

Zhiling Zheng, Oufan Zhang, Ha L Nguyen, Nakul Rampal, Ali H Alawadhi, Zichao Rong, Teresa Head-Gordon, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science*, 9(11):2161–2170, 2023.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.

Zexuan Zhong, Tao Lei, and Danqi Chen. Training language models with memory augmentation. *arXiv preprint arXiv:2205.12674*, 2022.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *ICLR 2023*, 2023.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are humanlevel prompt engineers. *arXiv preprint*, arXiv:2211.01910, 2022. Also see project Automatic Prompt Engineer (APE).

Yuhao Zhou, Yiheng Wang, Xuming He, et al. Scientists' first exam: Probing cognitive abilities of mllm via perception, understanding, and reasoning. *arXiv preprint arXiv:2506.10521*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, ZhangRen Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expertlevel medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Detail Illustration of Taxonomy for Scientific Agents

A.1 Elaboration between Scientific Agents and General-purpose agents

The workflow of all agents can be abstracted into three highly integrated stages: task, process, and result. The performance of different types of agents varies in these stages, specifically in aspects such as perception, cognition, decision-making, and action. For general-purpose agents, the workflow is more short-term, focusing on daily tasks that address a wide range of user demands. On the contrary, scientific agents are designed to follow the fixed processes of scientific research, with their workflow structured around six main stages: literature analysis, hypothesis generation, experiment design, experiment verification, result summarization, and feedback evaluation.

As illustrated in Table 7, differences in goal orientation result in significant disparities between general-purpose agents and scientific agents across multiple dimensions, from construction, design to evaluation.

Table 7: Multi-dimensional comparison between general-purpose agent and scientific agent.

Dimensions	General-purpose Agent	Scientific Agent
Knowledge Organization	Broad, flexible organizations	Domain-specific and high knowledge density
Knowledge Injection	Mostly relay on Pretraining	Modality-specific injection & complex context learning
Tool Integration	Diversities and utilies, easy to integrate	Specialized tools, hard to integrate
Memory	Task-specific and short-context retention	In-depth memory for long-term knowledge retention
Reasoning	General reasoning capabilities	Advanced reasoning for scientific discovery
Collaboration	Low collaboration requirements	Specialized role agents
Evaluation	User satisfaction and tasks diversity	Reproducibility, domain novelty and factuality.

A.2 Agent As Assistant

Initially, pre-trained models exhibited extensive capabilities in comprehending natural language broadly, though they were deficient in depth Lee et al. (2024b). Scientific agents at this juncture predominantly concentrated on an assistant-level role. Technically, they are designed to perform functions specific to distinct domain areas. primarily focusing on tasks such as answering questions, integrating knowledge, and providing partial assistance within specific domains.

Construction strategy. Assistant-level agents frequently utilize open-source models Bai et al. (2023); Touvron et al. (2023); Jiang et al. (2023) as the basemodel. These agents generally conduct knowledge organization and injection via fine-tuning through meticulous incorporation of domain-specific knowledge. Basic methods facilitate the effective domain knowledge injection via simple

Q&A tasks as BioGPT Luo et al. (2023), BioMedGPT Yang et al. (2023) ChemAU Li et al. (2025a), AstroLLaMA Nguyen et al. (2024) et al. Given the multimodal information of scientific data, NatureLM Xia et al. (2025) combines sequence information such as DNA, proteins into text to generate a training corpus. MolecularSTM Liu et al. (2022) employs contrastive learning techniques to effectively align CIF graph and sequence modalities. More recently, DeepSeek-Prover-V2 Ren et al. (2025b) and Ether0 Narayanan et al. (2025a) use Reinforcement Learning (RL) to train agents to answer domain complex questions through reasoning, like formal generation and chemical formula.

Capability scope. Assistant-level agents are generally constrained to executing specific tasks within particular research domains, lacking the capacity for seamless integration across the comprehensive research process. As illustrated in Fig. 2, the operational contexts for assistant agents primarily involve domain queries. They are frequently trained to conduct domain knowledge question answering, exemplified by instances such as DAARWIN 1.5 Xie et al. (2024), NatureLM Xia et al. (2025), MedAlpaca Han et al. (2023), or to perform domain-specific tasks like DeepSeekProver Ren et al. (2025c) or Ether0 Narayanan et al. (2025a) and HypoGen O'Neill et al. (2025). A minority of agents Thulke et al. (2024); O'Neill et al. (2025) may incorporate search tools or reasoning strategies to augment their response capabilities. However, they remain deficient in governing the entire scientific research trajectory. Notwithstanding, the pronounced specialization and compactness of assistant-level agents. Examples of such efficacy are observed with ChemBERTa Chithra et al. (2020) in chemical field applications and SciBERT Beltagy et al. (2019) in the extraction of scientific literature.

A.3 Agent As Partner

Partner-level agents acquire Agent-like attributes via systematic tool integration such as real-time data acquisition, advanced simulations, intuitive display, and traceable interactions. This progression is supported by improved reasoning, knowledge of pre-trained models, and tool integration.

Construction strategy. The primary emphasis of partner-level agents has transitioned from the development of models to the design of architectures. Closed-source, large-scale models exhibit capabilities that not only significantly exceed general performance metrics but also surpass human expert levels in benchmarks including ScienceQA Lu et al. (2022) and GPQA Rein et al. (2023). In conjunction with the reduction in costs, this shift has incentivized scientific agents to employ a "closed-source model + sufficient context" approach, thereby optimizing their potential for scientific discovery. By deconstructing the primary tasks, partner-level agents are equipped with relevant contextual descriptions for distinct subtasks, and employ either established or customized agent frameworks to develop workflows. Automated laboratory settings, like A-Lab Szymanski et al. (2023), Coscientist Boiko et al. (2023b), enable language models to execute directed experimental tasks through the organization of hardware interfaces or simulations. Conversely, more agents like ChemCrow Bran et al. (2023), Crispr-GPT Huang et al. (2024a), Organa Darvish et al. (2025). are dedicated to the integration of retrieval, analysis, and other tools to fulfill potential design objectives.

Capability scope. Most partner-level agents perform competently within specific domain topics, independently consulting literatureLála et al. (2023); Li et al. (2024f) and generating hypotheses Yang et al. (2024b) or experimental designsHuang et al. (2024a); Bran et al. (2023); Han et al. (2023). Some automated laboratories can also complete basic experiments. These agents often require highly constrained, customized scientific research environments, and many agents are limited to knowledge acquisition tools and do not integrate appropriate resources in experimental settings. Consequently, despite demonstrating robust design and analysis capabilities on fixed tasks, they often lack verifiability and data-driven characteristics Bran et al. (2023). Enhancements in reasoning, collaboration, and memory have, to some extent, improved the reliability of task outcomes; however, they are still unable to deal with domain subjects independently.

A.4 Agent As Avatar

Avatar-level intelligent agents can engage in the entire scientific research lifecycle through enhancements of targeted capability. They hold significant potential to promote scientific exploration, facilitate knowledge discovery, and expand the frontiers of science.

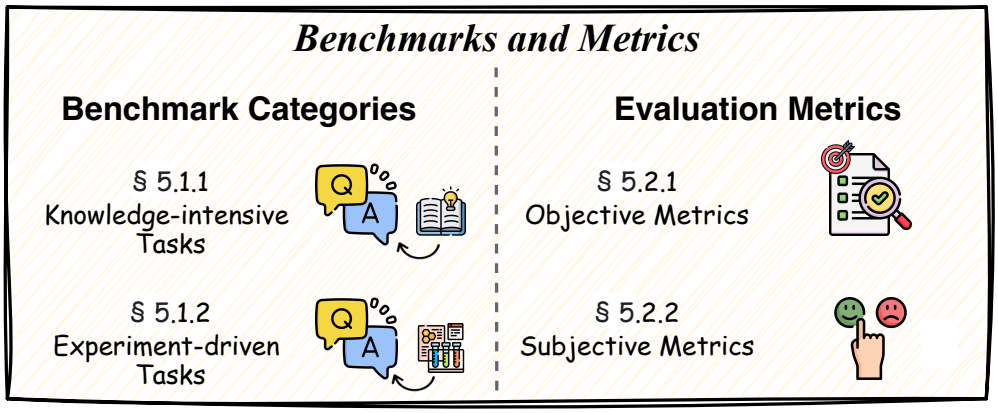


Figure 6: An overview of scientific agent benchmarks and evaluations metrics.

Construction strategy. Avatar agents are characterized by robust reasoning, deep memory, and strong collaboration. Alphaevolve Novikov et al. (2025) utilizes the original Gemini-2.5-pro to enhance endogenous reasoning, while Biomni Huang et al. (2025a) leverages reasoning for the orchestration of diverse toolchains. Robin Ghareeb et al. (2025) introduces a highly personalized approach to multi-agent collaboration. In this approach, Crow conducts a concise literature review to identify disease mechanisms and experimental strategies; Falcon performs a comprehensive evaluation of candidate therapeutic drugs; Finch autonomously analyzes experimental data from bioassays to facilitate new drug discovery. Additionally, AgentRxiv Schmidgall & Moor (2025) and Agent Laboratory Schmidgall et al. (2025) are built upon multi-agent collaboration. Furthermore, Alphaevolve Novikov et al. (2025), Biomni Huang et al. (2025a), Robin Ghareeb et al. (2025), and others are capable of enhancing memory through fusion based on prior interactions.

Capability scope. Studies have demonstrated that the capabilities of current agents are significantly limited by their orchestration frameworks Toledo et al. (2025). While Avatar-level agents are gradually broadening their capability scope in the scientific domain, their potential remains largely untapped. For instance, Biomni Huang et al. (2025a) excels in biological analysis and assists in multiple domains. Robin Ghareeb et al. (2025) can verify new drug discoveries within six months, and Alphaevolve Novikov et al. (2025)s exploration of computer algorithms underscores the considerable potential of avatars in scientific research. These agents can enhance the efficiency of scientific exploration at various research stages. However, due to the vast range of different fields, the breakthroughs achieved by current scientific agents have primarily been achieved in computer science and biochemistry. Expanding their applications into more specialized domains remains a challenging issue that warrants further investigation.

B Benchmarks and Metrics

A robust benchmark serves as a crucial tool for assessing the current state of a field and directing future advancements. As shown in Fig. 6, this section systematically reviews the benchmarks of scientific agents from two dimensions: benchmark categories and evaluation metrics.

B.1 Benchmark Categories

Table 8 summarizes existing benchmarks for scientific agents, categorizing into two categories: knowledge-intensive tasks and experiment-driven tasks, encompassing the entire lifecycle of scientific research.

B.1.1 Knowledge-intensive Tasks

Contrary to agents for general tasks, which require a substantial amount of common sense with an extensive knowledge domain, scientific agents are frequently confronted with complex, domain-

Table 8: Statistics of benchmarks for scientific agent evaluation. Application Stages is the research process stage in which the method is involved: L indicates Literature Mining, H indicates Research Hypothesis, D indicates Experiment Design, V indicates Experiment Verification, A indicates Analyst and Result, E indicates Evaluation and Review.

Category	Benchmark	Domain	Application Stages	Descriptions	Question Type	Links
Knowledge-intensive Tasks	BioMaze Zhao et al. (2025)	Biology	H D A	Complex Biological Systems Reasoning	True/False & Open-ended	Link
	Biokgbench Lin et al. (2024)	Biology & Medical	L A E	Hard Science QA	Short GT	Link
	Tomato-Chem Yang et al. (2024b)	Chemistry	L H E	Chemistry Hypothesis Discovery	Sequences GT	Link
	SurveyBench Yan et al. (2025)	Computer Science	L A	Survey Report Generation	Open-ended	Link
	MMLU-Pro Wang et al. (2024b)	General	L A	General Reasoning QA	Choices	Link
	HLE Hendrycks et al. (2025)	General	L A	Challenge Science QA	Short GT	Link
	ScienceQA Lu et al. (2022)	General	L A	Gereral Science QA	Choices	Link
	GPQA Rein et al. (2023)	General	L A	Hard Science QA	Hybrid GT	Link
	SuperGPQA Du et al. (2025b)	General	L A	Hard Science QA	Choices	Link
	SciMON Wang et al. (2023c)	General	L H	Scientific Hypothesis Generation	Open-ended	Link
	CiteBench Funkquist et al. (2022)	General	L A	Scientific Citation-Text Generation	Open-ended	Link
	SciGen Moosavi et al. (2021)	General	L A	Scientific Table Description Generation	Open-ended	Link
	ALCE Gao et al. (2023b)	General	L A	Scientific Citation-Text Generation	Open-ended	Link
	ReviewCritique Du et al. (2024a)	General	L E	Paper Review Generation	Hybrid GT	Link
	Reviewer2 Gao et al. (2024c)	General	L E	Paper Review Generation	Open-ended	Link
	DeepResearch Bench Du et al. (2025a)	General	L A	DeepResearch Reort Generation	Open-ended	Link
	Scientists' First Exam Zhou et al. (2025)	General	L A	General Science QA	Choices & Numeric	Link
	ArXivQA Li et al. (2024c)	Hybrid	L A	Multimodal Science QA	Choices	Link
	MR-Ben Zeng et al. (2024b)	Hybrid	L A	Science Reasoning QA	Choices	Link
	MMSci Li et al. (2024g)	Hybrid	L A	Multimodal Science QA	Choices	Link
	FigureQA Kahou et al. (2017)	Hybrid	L	Science Chart QA	Choices	Link
	DiscoveryBench Majumder et al. (2024)	Hybrid	H	Hypothesis Generation	Open-ended	Link
	SciEval Sun et al. (2024)	Hybrid	L H E	Scientific Research QA	Choices	Link
	MedXpertQA Zuo et al. (2025)	Medical	L A	Hard Medical Exam	Choices	Link
	MedQA Jin et al. (2021)	Medical	L A	Medical Exam	Choices	Link
	PharmaBench Niu et al. (2024)	Medical	L A V E	ADMET Property Prediction	Discrimination	Link
	LLM-SRBench Shojaee et al. (2025b)	Physics	A	Scientific Equation Discovery	Sequences GT	Link
Experiment-driven Tasks	LAB-Bench Laurent et al. (2024)	Biology	L D V A	Comprehensive Biology Tasks	Sequences GT	Link
	GenoTEX Liu & Wang (2024)	Biology	L D V A	Genomic Data Analysis Generation	Open-ended	Link
	MLE-Bench Chan et al. (2025)	Computer Science	D V A	Machine Learning Experiment Implementation	Sequences GT	Link
	MLAgentBench Huang et al. (2024b)	Computer Science	D V A	Machine Learning Experiment Implementation	Sequences GT	Link
	MASSW Zhang et al. (2025c)	Computer Science	L D A E	Scientific Publications Summarization	Open-ended	Link
	PaperBench Starace et al. (2025)	Computer Science	L D V A	Scientific Publications Replication	Open-ended	Link
	Exp-Bench Kon et al. (2025)	Computer Science	L H D A E	Scientific Publications Replication	Open-ended	Link
	DS-Bench Ouyang et al. (2025)	Data Science	D V A	Data Analysis	Choices & Sequences GT	Link
	SciClaimHunt Kumar et al. (2025)	General	D V E	Claim Verification	Hybrid GT	Link
	DiscoveryWorld Jansen et al. (2024)	Hybrid	H D V A	Scientific Experiment Implementation	Performance Scoring	Link
	ScienceAgentBench Chen et al. (2024b)	Hybrid	L D V A	End-to-end Scientific Tasks Impletation	Sequences GT	Link
	SciCode Roberts et al. (2024)	Hybrid	D V A	Code Generation	Sequences GT	Link
	ScienceBoard Sun et al. (2025b)	Hybrid	D V A	Scientific Workflow Implementations	Experiment Result	Link
	AgentClinic Schmidgall et al. (2024)	Hybrid	L D A	Clinical Decision Evaluation	Sequences GT	Link
	CORE-Bench Siegel et al. (2024)	Hybrid	L D V A	Scientific Publications Replication	Open-ended	Link
	SDRBench Zhao et al. (2021)	Hybrid	D V A	Scientific Data Reduction	Sequences GT	Link
	Auto-Bench Chen et al. (2025c)	Physics	D V A	Simulation-based Hardware Verification	Verilog Scripts	Link

specific tasks. Knowledge-intensive undertakings emphasize the agent's cognitive proficiency in domain-specific knowledge rather than general capabilities. This is evident in the production of reports on Literature mining, research hypothesis, experiment design, analysis and results, as well as evaluation and review stages, encompassing tasks associated with knowledge dissemination throughout the scientific research process.

Literature mining Organizing extensive literature or domain-specific corpora enables the generation of domain-related questions and answers, assessing the agent's deep-diving capabilities. This often requires logical reasoning. ScienceQA Lu et al. (2022), GPQA Rein et al. (2023), and SuperGPQA Du et al. (2025b) evaluate competencies across broad domains with a focus on reasoning abilities. HLE Hendrycks et al. (2025) addresses complex deep knowledge, while MedQA Jin et al. (2021), MedXpertQA Zuo et al. (2025), and related projects analyze specific domains.

Research hypothesis Valuable hypotheses are crucial for scientific progress and indicate an intelligent agent's understanding of a domain. Current hypothesis generation tasks are mainly divided into literature-based and experiment-based approaches. The literature-based method, used by SciMON Wang et al. (2023c) and Tomato-Chem Yang et al. (2024b), creates new hypotheses from historical documents, while the experiment-based method, as shown by DiscoveryBench Majumder et al. (2024), uses experimental results to find new patterns.

Analysis and result. The analysis is key to interpreting research results, requiring scientific knowledge for a credible explanation. Benchmarks in this process are crucial for scientific inquiry Li et al. (2024c); Zeng et al. (2024b); Wang et al. (2024b); Jin et al. (2021); Zuo et al. (2025); Rein et al. (2023); Du et al. (2025b); Hendrycks et al. (2025); Shojaee et al. (2025b). After analysis, a struc-

tured report using systematic methods is essential. Structured knowledge greatly refines insights. CiteBench Funkquist et al. (2022) and ALCE Gao et al. (2023b) focus on factual citations, while DeepResearch Bench Du et al. (2025a), SurveyBench Yan et al. (2025) assess research reports.

Evaluation and review. Scientific agents have revolutionized the traditional evaluation and review process. Existing benchmarks, Reviewer2 Gao et al. (2024c), ReviewCritique Du et al. (2024a), focus on measuring agents' alignment with human reviewers.

B.1.2 Experiment-driven Tasks

Experiment-driven tasks thoroughly incorporate the assessment of the scientific agent's ability to utilize tools. They particularly emphasize evaluating the agent's proficiency in employing tools for the execution of experiment design and validation, or for undertaking multi-process explorations rooted in scientific inquiries.

Experiment design and verification. Experiment-driven tasks aim to assess an agent's capability to accomplish objectives autonomously, without reliance on expert intervention. Various benchmarks across distinct domains, such as DS-Bench Ouyang et al. (2025), MLAgentBench Huang et al. (2024b), ScienceBoard Sun et al. (2025b), and SDRbench Zhao et al. (2021), among others, evaluate the extent to which experiments devised by the agent are completed.

Multi-process exploration. Some agents are set to automatically explore based on certain specific topics. DiscoveryWorld Jansen et al. (2024) explores scientific research from a fixed scenario, and ScienceAgentBenh Chen et al. (2024b), Paperbench Starace et al. (2025), CORE-Bench Siegel et al. (2024) measure the research capabilities of scientific agents based on literature reproduction.

B.2 Evaluation Metrics

Various evaluation metrics have been devised to quantify the behavioural quality and performance limits of scientific agents throughout the whole stages of the research process. Depending on the nature of the underlying measurement, these core metrics are typically classified into two categories: objective evaluation metrics and subjective evaluations.

B.2.1 Objective Metrics

Based on verified result. Most benchmarks incorporate question-answering tasks into multiple-choice questions framed with objective responses or sequential answers. In the context of standard multiple-choice questions, basic `accuracy` serves as the metric for evaluation. For complex reasoning questions with sequential answers or verified results, metrics such as `pass@k` and `cons@k` are frequently employed. The `pass@k` metric denotes the likelihood that the agent will attempt the problem k times, where c indicates the number of correct samples in total samples n and $\mathbb{E}_c$ refers to the expectation of c,

$$\text{pass@}k := \mathbb{E}_c\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right], \tag{1}$$

`cons@k` measures the probability of achieving consistency over k attempts by the agent via majority vote,

$$\begin{aligned} \text{cons@}K &:= \frac{1}{n}\sum_{i=1}^{n} 1[\hat{a}_i = y_i], \\ where\ \hat{a}_i &= \arg\max_a \sum_{j=1}^{K} 1[y_{\{i,j\}} = a], \end{aligned} \tag{2}$$

For a specific domain, such as the effectiveness of synthetic molecules, differences in predicted properties, etc., can all be used as verifiable objective evaluation criteria.

Based on trajectory. In addition to the success rate of individual tasks, agent process indicators are often incorporated to provide a more comprehensive evaluation. For tasks with distinct steps, the completion of sub-goals is assessed after breaking down the process Xu et al. (2024). To evaluate

the trajectory effectively, precision and recall can be calculated for each step Liu et al. (2023c). The efficiency of the agent's process can be measured by the average completed steps Liu et al. (2023c).

Based on output similarity. Similarity-based metrics provide a foundation for evaluating open-ended question-answering systems. The primary methodologies employed include BLEU Papineni et al. (2002), BertScore Zhang et al. (2020), and Semantic Textual Similarity (STS) Majumder et al. (2016). BLEU evaluates the overall consistency between the generated text and the reference text by examining n-gram word segment overlap, necessitating that the reference text possess a high degree of semantic structure standardization. BertScore assesses the maximum cosine similarity of contextual embeddings by leveraging the Bert model across various domains. STS evaluates similarity with the reference answer at the sentence level.

Based on factuality. Robust factuality assessment underpins the reliability of scientific-agent outputs. Metrics like ANAH-v2 Gu et al. (2024), FactScore Min et al. (2023), LongFact Wei et al. (2024)all use LLM-as-judge to establish a rigorous workflow to evaluate the factuality of each proposition on sentence-level atom answer. Although this pipeline yields fine-grained, documentable judgments, LLM-as-judge prompts reduce manual and computational overheadtrading methodological transparency for efficiency.

B.2.2 Subjective Metrics

Research activities like hypothesis formulation, experimental design, data analysis, and synthesis in research report creation, along with evaluation and review, are often subjective tasks. The prevailing evaluation methodologies frequently incorporate an approach that combines granular evaluation decomposition with the deployment of LLM-as-judge Gu et al. (2025). The assessment criteria can be accurately evaluated by disaggregating and quantifying the measurement components of the produced object. For instance, Deepresearch Bench evaluates the generated report across four dimensions: comprehensiveness, insight, adherence to instructions, and readability. Shin et al. (2025) utilizes clarity, validity, novelty, and impact as the dimensions for decomposing the report for scoring purposes. Concurrently, to enhance the reliability of LLM-as-judge, manual evaluation is frequently incorporated as a reference, or reference samples are utilized as prompts. Simultaneously, the DeepResearch Bench Du et al. (2025a) evaluates and normalizes both the reference samples and generated samples, further quantifying the scores in terms of relative degrees.

C Challenge and Prospect

Existing research is still far from the expected scientific agent prospect, and still faces challenges in many dimensions. This section will discuss the challenges and promising research directions of scientific research, aiming to further promote the development of the field.

Factualism and rationality. Evaluating the rationality and factual accuracy of scientific experimental designs still remain challenge for existing scientific agents. While a controlled degree of creative hallucinations can potentially foster innovation in scientific discovery Abdel-Rehim et al. (2025), LLMs possess a pronounced tendency to generate fabricated scientific arguments. Furthermore, In addition, wet experiments in natural sciences often require a lot of resources, which makes it more complicated to align scientific agents with verifiable facts. In the future, this can be alleviated through further tool integration, knowledge verification mechanisms, and iterative feedback reasoning, so as to guide data-driven methods with scientific soft literacy.

Framework adapted complex scientific tasks. Scientific agents remain relatively lagging compared to other types of agents. Many existing systems redundantly reinventing wheels across various domains and lacking cohesive, well-structured architectures. Scientific agents such as Biomni Huang et al. (2025a)) support complex workflows by integrating various biological tools, but they do not adopt existing agent frameworks because most agent frameworks Chase (2022); LangChain Inc. (2024); Wu et al. (2023); Lu et al. (2025); Moura & Contributors (2024) have high code complexity and are difficult to integrate. Designing flexible agent frameworks tailored to the specificities of scientific research tasks, such as Octotools' tool card design Lu et al. (2025), to meet the research needs of diverse fields, may be a promising direction for future exploration.

Self-improvement iteration. With rapid technological advancement, self-iterative progress is essential for the long-term development of scientific agents. During self-iteration, memory is continuously accumulated as a communication unit, with episodic memory gradually internalized Pink et al. (2025). More autonomous approaches employ self-reflection iterations to develop a refined experiential insight library. The benefits of continual learning in mitigating catastrophic forgetting offer valuable guidance for integrating parameter memory into LLMs Guo et al. (2025). Future research should determine how to optimally balance episodic and parameter memory in agents to support continuous evolution.

Interaction optimization for scientific exploration. Optimizing the interactions among scientific agents during autonomous exploration, particularly in human-machine and multi-agent collaboration, is crucial for advancing research. Current multi-agent systems are often built on fixed, general models, which limits their diversity and hinders their development. Expanding to a wider range of models or introducing human collaboration can alleviate this problem to some extent. The Robin system (Ghareeb et al. (2025)) provides a powerful framework with a clear division of labor, demonstrating the superiority of interaction design. Future paradigms that combine general and specialized models, and foster the interaction of artificial intelligence and human expertise, hold great potential for enhancing the capabilities of scientific agents and provide fertile ground for future research.

Multi-disciplinary agent. Disciplines are not entirely isolated entities. While different fields exhibit distinct cognitive styles and reasoning preferences, they simultaneously share certain common elements. This facilitates the implicit transfer of knowledge across various disciplines. In a similar manner to NatureLM Xia et al. (2025) and Biomni Huang et al. (2025a), which integrates sequential information such as proteins, DNA, and molecules for training purposes, resulting in performance enhancements across multiple domains, scientific agents might augment their professional expertise by strengthening their capacities across related disciplines.

Scientific evaluation and verification The extant benchmarks are driven by tasks, rendering segments of the research process into problems that can be empirically verified Sun et al. (2025b); Siegel et al. (2024); Chen et al. (2024b); Roberts et al. (2024); Jansen et al. (2024). Nonetheless, a disparity exists between scientific research tasks in open-ended scenarios and the corresponding evaluations. Such evaluation of scientific research tasks can be considered an undecidable problem. Identifying appropriate relaxation conditions to conceptualize it within the NP class, or employing approximate reductions such as probabilistic or randomization, remains an area worthy of further investigation. AI-driven research still must adhere to the falsifiability and reproducibility inherent in scientific methodologies.