Exploring Agentic Multimodal Large Language Models: A Survey for AIScientists

Jinglin Jian [†‡] **Yi R. Fung** [†] **Denghui Zhang** [◇] **Yiqian Liang** [§]
Qingyu Chen [♡] **Zhiyong Lu** [¶] **Qingyun Wang** [*]
[†]University of Illinois at Urbana-Champaign [§]University of Cambridge
[♡]Yale University [◇]Stevens Institute of Technology [*]William & Mary
[¶]National Library of Medicine, National Institutes of Health
jjian@scripps.edu, {jj50,yifung2}@illinois.edu, dzhang42@stevens.edu
yl992@cam.ac.uk, qingyu.chen@yale.edu, zhiyong.lu@nih.gov, qwang16@wm.edu

Abstract

Agentic Multimodal Large Language Models (MLLMs) are reshaping scientific discovery. They do more than answer questions: they read multimodal evidence, reason over scientific constraints, use tools, and close parts of the research loop. This survey explains how such systems are built for scientific discovery, moving from multimodal perception to training, inference, evaluation, and human-AI collaboration. Across this pipeline, scientific agents face challenges that standard assistant tasks may not reveal, including scarce expert annotations, evolving domain knowledge, complex reasoning, limited tool interfaces, safety and toxicity risks, and results unsupported by evidence. We highlight design choices for turning agentic MLLMs into trustworthy, evidence-grounded collaborators. Curated resources are available at https://github.com/JianJinglin/awesome-agentic-AIScientists.

1 Introduction

A golden era for scientific discovery is upon us. Remarkable progress in large language models and multimodal foundation models has unlocked transformative AI systems capable of reasoning and generating human-like outputs across various tasks and domains. The 2024 Nobel Prize in Chemistry[1] further highlights the impact of computational methods on scientific discovery. Scientists increasingly use agentic AI systems to help tackle research problems, including designing new drugs (M. Bran et al., 2024) and materials (Szymanski et al., 2023). This recognition signals a paradigm shift, catalyzing the rapid emergence of AIScientist systems (Figure 1).

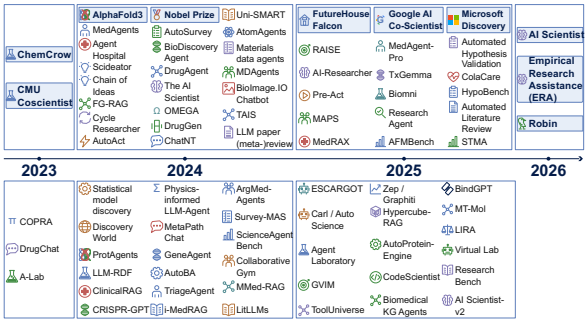


Figure 1: Evolution of Major AIScientist Systems

The concept of the "AIScientist," first introduced by Lu et al. (2024a), envisions a framework for fully automated scientific discovery. Powered by agentic multimodal large language models (MLLMs), these systems aim to support the entire research lifecycle: understanding papers (Radensky et al., 2024; Wang et al., 2024c), generating new hypotheses (Wang et al., 2024a; Li et al., 2024e; Baek et al., 2025), planning and conducting experiments (Boiko et al., 2023; Tom et al., 2024), analyzing results (Jansen et al., 2025; Roohani et al., 2024), drafting manuscripts (Wang et al., 2024b), and reviewing papers (Du et al., 2024; Weng et al., 2024). Recent Nature studies show this transition from assistant tools to scientific agents in practice, enabling domain-expert-level data analysis, in vitro experimental validation, and expert-level empirical software generation (Gottweis et al., 2026; Ghareeb et al., 2026; Aygün et al., 2026). These systems show that scientific autonomy depends on verification, domain knowledge, and human oversight throughout the research process.

This survey provides a systematic account of how an agentic MLLM-based scientist can be designed to support scientific work. Rather than treating scientific assistants as generic chatbots with tools, we frame the central design question as how multimodal perception, memory, planning, action,

[‡]Current affiliation: Scripps Research.

[1]https://www.nobelprize.org/prizes/chemistry/2024/press-release/

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1089–1133
November 6-10, 2026 ©2026 Association for Computational Linguistics

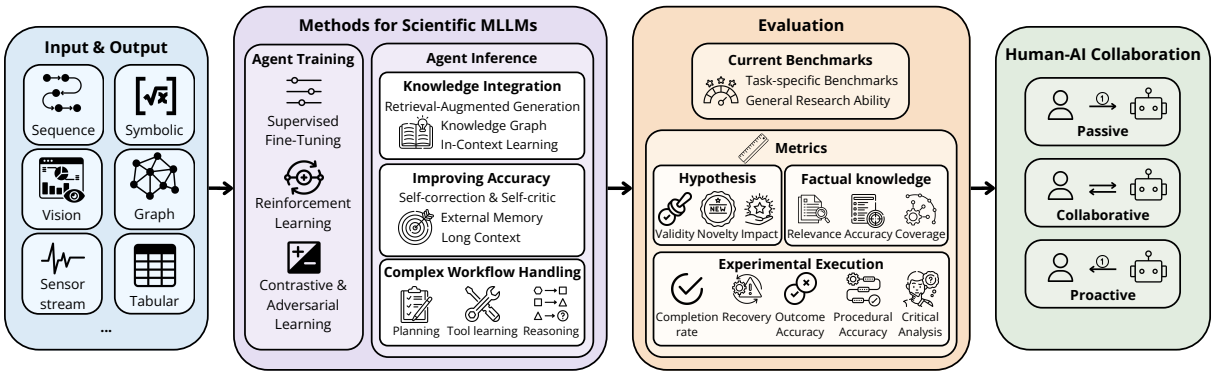


Figure 2: **Overview of the agentic MLLM framework for scientific discovery.** The pipeline includes **Input & Output** modalities (Sec 3), **Methods** covering agent training and inference (Sec 4), **Evaluation** with benchmarks and metrics (Sec 5), and **Human–AI Collaboration** (Sec 6).

and verification should be composed into a closed-loop research system.

Progress also depends on benchmarks grounded in concrete scientific tasks and interfaces that let scientists delegate work, inspect evidence, and correct agents as a study develops. This survey provides a pipeline framework for developing agentic MLLMs in scientific contexts (Figure 2), spanning multimodal inputs and outputs, agent training and inference, evaluation, and human-AI collaboration.

2 Scientific Workflow and the Need for Multimodal Agentic AI Scientists

Scientific discovery remains a fundamentally human pursuit based on iterative inquiry, but modern research has become increasingly complex. Today's scientific workflows connect literature, experimental measurements, and computational results across repeated cycles of inquiry. The need to reason across these sources motivates the development of the **Agentic AIScientist**. We view such agents as modular systems that integrate Learning, Memory, Action, Cognition, and interaction with the environment (Durante et al., 2024). The development of agentic AIScientists can help automate time-consuming tasks, enabling researchers to focus on creative challenges. The details of the agent's composition and scientific process are provided in Appendix B.

Operational Scope An **agentic AIScientist** is an MLLM-centered system that can understand heterogeneous scientific data, such as text, images, molecules, equations, code, and experimental measurements, and coordinate actions toward a scientific objective. Beyond perception, it tracks study progress, plans steps, selects tools or collaborators, checks results, and revises plans to produce reproducible scientific artifacts. This workflow-level framing (Figure 3) distinguishes it from a *tool-augmented MLLM*, which may call external tools for a single response but does not necessarily maintain closed-loop autonomy over a research workflow.

3 Multimodal Perception

Scientific domains differ from generic multimodal settings because claims are judged against specialized objects, measurement protocols, and expert evidence. Biomedical text, omics, molecules, materials, microscopy, climate, mathematics, and closed-loop discovery workflows all require such constraints (Beltagy et al., 2019; Gu et al., 2022; Chen et al., 2024e; Jansen et al., 2024). Following Si, Yang, and Hashimoto's expert-evaluation setup (Si et al., 2024), we treat a scientific domain as one where novelty, feasibility, and evidence support require community expertise. We use the definition of an agent as an interactive system that interprets language, visual inputs, and observations from its environment to produce actions (Durante et al., 2024); Figure 3 shows how these components support scientific work.

Perception must therefore preserve the scientific meaning of each input. It must preserve sequences (Wang et al., 2019b; Luo et al., 2023c), graphs (Gilmer et al., 2017; Wang et al., 2023f), images (Li et al., 2024h; Burgess et al., 2025; Alampara et al., 2025), instrument measurements (Barman et al., 2025; Pyzer-Knapp et al., 2025), and equations (Yang et al., 2024b; Cherian et al., 2024), while retaining units and provenance so errors do

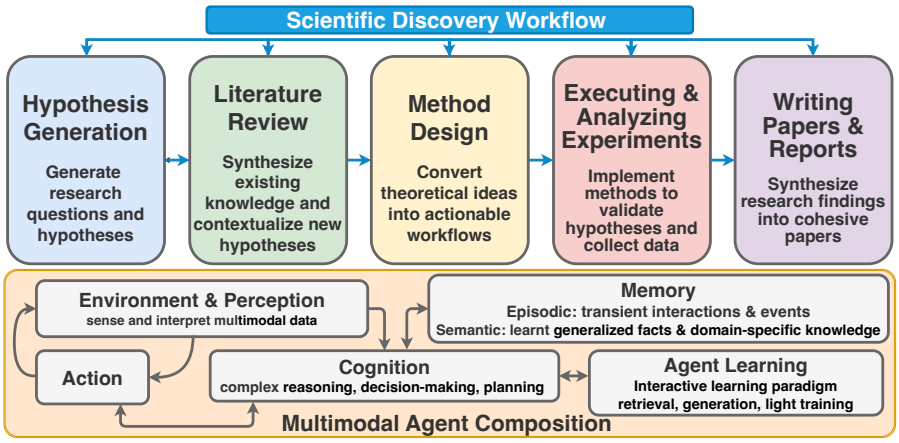


Figure 3: The Modern Scientific Discovery Workflow and Modular Architecture of the Agentic AI System.

not propagate into plans, tool calls, and conclusions (Wang et al., 2023b; Acosta et al., 2022).

3.1 Data Modalities

Sequence Sequence models cover scientific text (de Haan et al., 2025; Bolton et al., 2024; Wu et al., 2024), molecules (Ahmad et al., 2022; Wang et al., 2019b; Luo et al., 2023c), genomes (Fishman et al., 2025), and proteins (Jumper et al., 2021; Elnaggar et al., 2022; Lv et al., 2026). They capture long-range patterns in language and biological sequences (Lin et al., 2023; Kim et al., 2025b), but may miss geometry and physical constraints needed for molecular reasoning (Luo et al., 2023c; Madani et al., 2020).

Symbolic Symbolic inputs express proofs, laws, algorithms, code, and simulations (Yang et al., 2024b; Cherian et al., 2024; Liu et al., 2024c; Majumder et al., 2024; Tang et al., 2024a). GPT-f and LeanCopilot generate proofs checked in Metamath or Lean (Polu and Sutskever, 2020; Song et al., 2024; Thakur et al., 2023), and symbolic agents can control scientific software (Schmidgall et al., 2025; Li et al., 2025d; Song et al., 2023); software can check their outputs, although exact syntax and translating informal context remain challenging.

Vision Vision models handle medicine, microscopy, geoscience, and scientific figures (Fallahpour et al., 2025; Li et al., 2024h; Burgess et al., 2025; Yan et al., 2024). CLIP/BLIP-style alignment and diffusion models support image-text grounding, molecule design, protein-ligand modeling, and segmentation (Radford et al., 2021; Li et al., 2022; Zhang et al., 2025c,a; Guan et al., 2024; Qin et al., 2025), but fine-grained reasoning over noisy scientific images remains difficult (Chen et al., 2024a; Lei et al., 2024b; Alampara et al., 2025). Retaining source images allows later checks of extracted features.

Graph Graphs encode relations in biology, chemistry, and physics (Jumper et al., 2021; Yi et al., 2022; Gilmer et al., 2017; Reiser et al., 2022; Sanchez-Gonzalez et al., 2020). Molecular and knowledge graphs provide structured priors for molecule or protein reasoning (Liu et al., 2024d; Luo et al., 2023c; Wang et al., 2023f; Xu et al., 2023; Liang et al., 2023b; Abdine et al., 2024), but changes in entities and relations create maintenance costs (Feng et al., 2024; Zeng et al., 2019).

Sensor Stream Sensor streams support biomedicine, experimental physics, climate, and spatiotemporal monitoring (Barman et al., 2025; Pyzer-Knapp et al., 2025; Liang et al., 2024; Cong et al., 2022). NormWear aligns physiological signals with language (Luo et al., 2024), while ClimaX and DiffSTG model climate or spatiotemporal dynamics (Nguyen et al., 2023; Wen et al., 2023a); however, sparse, noisy measurements remain difficult to align across instruments (Zhang et al., 2024c). Sampling and calibration records help distinguish experimental changes from measurement artifacts.

Other Experimental Data Scientific research also relies on structured instrument outputs such as tabular, single-cell, and spatial omics (Wu and Xie, 2025; Yang et al., 2022; Wen et al., 2023c). Geneformer, scGPT, CellPLM, and Nicheformer model these data (Theodoris et al., 2023; Cui et al., 2024; Wen et al., 2023b; Schaar et al., 2024), but high dimensionality, batch effects, and inconsistent formats limit transfer (Hao et al., 2024; Heimberg et al., 2023; Rosen et al., 2023). Recording pre-

processing choices supports comparisons across studies.

3.2 Multi-modal Knowledge Fusion

Agentic MLLMs integrate modalities through three fusion strategies (Tu et al., 2024; Schouten et al., 2024). **Intermediate Fusion** encodes modalities separately and merges them by concatenation, attention (Xu et al., 2023), or latent alignment (Zhang et al., 2025c); it supports interaction between modalities but requires careful alignment and calibration. **Late Fusion** combines independent model outputs (Krones et al., 2025); it is modular but weak at detecting cross-modal contradictions. **Early Fusion** combines raw inputs before encoding, but is rare because they require strict spatial or semantic registration. Fused representations should retain source information so disagreements between modalities can be traced to the observations.

Takeaway: Scientific perception must preserve the constraints and evidence that agents use to plan and act. **Sequences** miss geometry, **symbols** require exact syntax, **vision** is noisy, **graphs** evolve, **sensors** drift, and **omics** needs batch correction. **Intermediate fusion** is most useful when cross-modal evidence must interact; **late fusion** permits independent encoders and missing modalities, but requires an explicit cross-check when their predictions disagree.

4 Methods for Scientific MLLM Agents

Scientific Challenge Framing Frontier model families such as Claude and Gemini (Anthropic, 2024; Gemini Team et al., 2023) still struggle when scientific tasks require domain-specific reasoning, software use, and verifiable action (Phan et al., 2025; Sawamura et al., 2025). The difficulty comes from three sources: expert annotations are scarce and costly (Beltagy et al., 2019); scientific reasoning depends on formulas, temporal dynamics, hypotheses, priors, and boundary conditions; and many tools or instrument interfaces were designed for human operators rather than autonomous agents (Chen et al., 2024e; Mandal et al., 2025). Methods for scientific MLLM agents must therefore link actions to evidence, update representations as knowledge changes, and use execution failures to revise the plan.

4.1 Agent Training

Training helps agents learn domain knowledge, tool use, and sequential decisions. In science, the bottleneck is not generic instruction following, but reliable supervision: expert labels are expensive, perturbations must remain scientifically valid, rewards need verifiers, and models must stay aligned with changing structures, protocols, and databases (Beltagy et al., 2019; Gu et al., 2022; Patil et al., 2024; Qin et al., 2024).

4.1.1 Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) adapts foundation MLLMs with curated input-output examples (Maleki et al., 2024; Wei et al., 2021). It supports molecular property prediction (de Almeida et al., 2025), protein design (Lv et al., 2026), therapeutic reasoning (Wang et al., 2025a), and API/-tool use through datasets such as Gorilla and ToolLLM (Patil et al., 2024; Qin et al., 2024). Full SFT updates all model weights; LoRA trains low-rank adapters while freezing the base model (Ma, 2025). MMed-RAG applies LoRA to biomedical VQA (Xia et al., 2024), but factual claims still need independent checks (Belouadi et al., 2023).

4.1.2 Reinforcement Learning

Reinforcement Learning (RL) learns action policies from rewards, supporting scientific tasks with sequential decisions, retries, or environmental feedback (Ahmed and Mohammed, 2024; AlphaProof and AlphaGeometry teams, 2024).

Preference Feedback Reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) align models with preference judgments in open-ended tasks (Ouyang et al., 2022; OpenAI, 2022; Lambert et al., 2023). Unlike standard RLHF, DPO learns from preferred and rejected responses without fitting a separate reward model (Rafailov et al., 2023). MAVIS combines automatically generated visual mathematics data with instruction tuning and DPO (Zhang et al., 2024e; Wei et al., 2022). ResearchAgent uses human-aligned reviewing models for iterative idea refinement at inference time (Baek et al., 2025); it does not report a new RLHF training procedure. However, collecting expert preferences is costly and cannot replace scientific validation.

Simulation & Verified Feedback Verifiers and simulators provide denser rewards when correctness is executable (Zhou et al., 2023a; Jha et al., 2024). Leanabell-Prover-V2 uses symbolic proof checks (Ji et al., 2025); AgentHospital and AtomAgents simulate scientific environments (Li et al., 2024c; Ghafarollahi and Buehler, 2025); and BindGPT or DrugGen use molecular scores as re-

wards (Zholus et al., 2025; Sheikholeslami et al., 2024). These rewards are automatically checkable, but depend on verifier accuracy and simulator fidelity (Zhang et al., 2025e).

4.1.3 Contrastive & Adversarial Learning

With scarce labels, contrastive learning (CL) brings related examples closer in representation space and separates unrelated examples (Zhou et al., 2023b; Gao et al., 2023a). Hard negatives and counterfactual pairs can improve generalization (Chen et al., 2024c; Roschewitz et al., 2025), while adversarial perturbations improve robustness for similar molecules or noisy measurements (Lan et al., 2024; Li et al., 2024a). However, poor augmentations or negative samples can erase meaningful scientific differences and destabilize training (Xiao et al., 2025; Salimans et al., 2016).

Comparative Analysis SFT, RL, and CL solve different bottlenecks. SFT is useful when expert examples are available; CL uses relationships between examples when labels are scarce; RL supports sequential decisions when validators or simulators can score actions. Their cost profiles differ as well: SFT/CL spend GPU hours upfront, parameter-efficient fine-tuning (PEFT) reduces that burden (Ma, 2025), and RL is usually heavier because rollouts, simulators, verifiers, and retries multiply compute (Zhang et al., 2025e).

The choice also depends on the modality: image–text pairs can support contrastive alignment, whereas molecular augmentations must preserve chemically relevant structure and omics negatives must account for batch effects. Retrieval is preferable to repeated fine-tuning when knowledge changes faster than reliable labels can be collected; RL becomes attractive when the required constraint has a trustworthy executable check. The choice depends on evidence, feedback reliability, and validation cost.

4.2 Agent Inference

Inference-time methods improve agents without changing weights (Snell et al., 2024). For scientific tasks, much of the cost then comes from retrieval, long contexts, tool calls, and validation (Hu et al., 2025a; Chen et al., 2024e). MCP-style protocols and tool ecosystems are emerging (Hou et al., 2025; Gao et al., 2025; Huang et al., 2025b), but many scientific APIs, datasets, and instruments remain closed, underspecified, or built for humans rather than agents.

4.2.1 Integrating Domain Knowledge

Retrieval-Augmented Generation (RAG) RAG grounds outputs in retrieved evidence. Biomedical agents retrieve guidelines or literature for clinical alignment and protein engineering (Wang et al., 2025c,b; Lu et al., 2024b; Liu et al., 2025c); real-time and task-specific variants add query refinement, structured prompts, or hypercube search (Lu et al., 2024e; Ghafarollahi and Buehler, 2024; Lei et al., 2024b; Xiong et al., 2024; Liu et al., 2024e; Shi et al., 2025). However, poor corpus quality or coverage can yield outdated evidence or omit competing findings. Retaining source passages makes claims easier to inspect.

Knowledge Graphs (KG) KGs impose structured constraints for proposal generation and executable graph querying (Selinger et al., 2024; Aamer et al., 2025; Matsumoto et al., 2025). RAG typically retrieves supporting passages; KGs make entity relationships explicit, allowing queries to follow links and check relational constraints. ESCARGOT queries a scientific graph through code during reasoning (Matsumoto et al., 2025). However, changing entities, relations, and confidence scores complicate maintenance.

In-Context Learning (ICL) ICL adapts behavior through demonstrations rather than weight updates (Brown et al., 2020; Dong et al., 2022). It works across symbolic proofs, geospatial analysis, medical segmentation, and protein/molecule tasks (Kasaura et al., 2025; Cherian et al., 2024; Singh et al., 2024b; Mansourian and Oucheikh, 2024; Kumar et al., 2024; Ma et al., 2025; Li et al., 2025d; Cai et al., 2024b). However, performance depends on demonstration relevance, order, and format; more examples increase context length and inference cost (Liu et al., 2021a; Fung et al., 2025; Liu et al., 2024b; Hu et al., 2025a; Bertsch et al., 2024).

4.2.2 Complex Workflow Handling

Planning ReAct, Pre-Act, and AutoAct interleave reasoning and actions (Yao et al., 2022; Rawat et al., 2025; Qiao et al., 2024). Scientific planners such as RAISE and LLM-RDF decompose tasks into logical queries, Bayesian optimization, or robotic actions (Oh et al., 2025; Ruan et al., 2024); Monte Carlo tree search (MCTS) and biomedical planners explore alternatives but add latency (Zhang et al., 2024a; Liu et al., 2024e).

Intermediate checks help catch failures before execution.

Tool Learning Scientific agents call databases, visual models, laboratory software, and robotic instruments (Fallahpour et al., 2025; McNaughton et al., 2024; Huang et al., 2024b; M. Bran et al., 2024; Roohani et al., 2024; Royer, 2024). ToolUniverse and Biomni show how standardized tool ecosystems can broaden access (Gao et al., 2025; Huang et al., 2025b), but laboratory automation requires documentation, permissions, and machine-readable interfaces (Boiko et al., 2023; Hou et al., 2025). Logging arguments, results, and errors supports reproducibility and recovery.

Multi-Step Reasoning Chain-of-thought and role-specialized agents improve step-wise inference (Chen et al., 2025b; Huang et al., 2024b). MT-Mol uses RDKit for chemical reasoning (Kim et al., 2025a), MedAgent-Pro improves explainability (Wang et al., 2025c), and iterative verification supports material property prediction (Ansari and Moosavi, 2024). Full-loop systems extend this pattern: AI Scientist variants use tree search for idea generation, coding, experiments, and writing (Lu et al., 2026; Yamada et al., 2025); ERA optimizes empirical software (Aygün et al., 2026); and Co-Scientist or Robin connect hypothesis generation with biomedical validation (Gottweis et al., 2026; Ghareeb et al., 2026). However, early errors can propagate, requiring intermediate checks.

4.2.3 Improving Accuracy

Self-Critique Self-correction lets agents evaluate and revise their own outputs (Shinn et al., 2023). ArgMed Agents use self-argumentation for clinical explainability (Hong et al., 2024), and GeneAgent reduces hallucinations by querying databases and analyzing data (Wang et al., 2024c). However, revisions add inference cost and require new evidence or external checks.

External Memory Memory preserves study context and tool traces (Ghafarollahi and Buehler, 2025). Episodic memory records actions and feedback; semantic memory stores general knowledge (Durante et al., 2024). Temporal-aware systems such as Graphiti and STMA add time sensitivity (Rasmussen et al., 2025; Lei et al., 2025), but retrieval becomes noisy as histories grow.

4.3 Multi-Agent System

Multi-agent systems decompose scientific work into roles such as literature search, protocol design, data analysis, safety review, and validation (Pandey et al., 2024; Tran et al., 2025). Debate, confidence-weighted voting, dynamic recruitment, and role assignment can improve coverage (Liang et al., 2023a; Chen et al., 2023; Tang et al., 2024b; AL-Mutairi et al., 2024; Kim et al., 2024; Wang et al., 2025b). Complementary roles can contribute different evidence or tools; however, agents with shared evidence may repeat errors at added cost.

Choosing a Workflow Use RAG for missing evidence, KGs for relation constraints, and branching search when reliable checks can screen plans before costly execution. Self-critique needs new evidence; multiple agents need complementary tools or independent checks. Tables 4 and 5 compare costs, including tokens, tools, retries, and expert time.

Takeaway: Choose scientific agent methods according to the available evidence and the checks that make their outputs reliable. **SFT** transfers expert patterns but is label-limited; **RL** is strongest with simulator or validator rewards but is expensive; **CL** helps low-label domains only when augmentations stay scientifically valid. At inference, **RAG/KG/ICL** ground actions in evidence, while **planning, tools, memory, and multi-agent roles** are useful only when they expose provenance, failure points, and independent checks.

5 Benchmarks and Evaluation

5.1 Current Benchmarks

Benchmark tables cover biological omics data analysis (Appendix Table 7), chemistry and materials (Appendix Table 8), physics (Appendix Table 9), mathematics (Appendix Table 10), environmental science (Appendix Table 11), geoscience (Appendix Table 12), multidisciplinary knowledge (Appendix Table 13), and general scientific discovery (Appendix Table 14). Existing benchmarks test capabilities such as multimodal question answering (ScienceQA (Lu et al., 2022)) and scientific code generation (SciCode (Tian et al., 2024)), making their task definitions important when comparing results. DiscoveryWorld (Jansen et al., 2024) evaluates multi-step discovery workflows.

Table 1 illustrates the range of biology benchmarks, from question answering to protocol reasoning and single-cell analysis. Appendix E.1 gives the full catalog, including task sources, modalities, test scales, and evaluation metrics.

Benchmark	Task	Modality					SC	#	MC
		Sq.	Sy.	V.	G.	O.			
PGxQA (Keat et al., 2024)	QA	✓					PubL., Man.	110k	OA., HL., Man., Sim., Suc.
MicroVQA (Burgess et al., 2025)	QA	✓		✓			Man.	1k	OA.
BioLLMBench (Sarwal et al., 2023)	BI.	✓	✓				Lit.	36	OA., Sim., Suc., Cost, Simp.
BioCoder (Tang et al., 2024a)	Code	✓	✓				Git., Web.	70k	OA.
BioProBench (Liu et al., 2025d)	Reas.	✓					PubL., Syn.	556k	OA., Err., Sim.
SimBench (Cao et al., 2021)	ScA.					✓T.	Lit.	8 cat.	OA., Sim.
DUD-E (Mysinger et al., 2012)	MD.	✓			✓	✓3D	PubL., Syn.	102	OA.
BaisBench (Luo et al., 2025a)	Omics	✓				✓T.	Lit., Syn.	229	OA.

Table 1: **A partial list of benchmarks for scientific MLLM agents (Biology).** The full biology table is provided in Appendix Table 7. Appendix E provides the benchmark catalog and detailed metrics across 10 scientific disciplines. *Note.* SC=task sources; #=test scale; MC=metrics. OA.=outcome accuracy; Err.=prediction error; Sim.=similarity; HL.=hallucination; Suc.=success rate; Simp.=simplicity; Man.=human annotation/evaluation. DUD-E: Sq.=SMILES; G.=atom–bond connectivity; 3D=structures. Other abbreviations are in Appendix E.

5.2 Metrics across Research Stages

As scientific agents take on more of a workflow, evaluation should distinguish the quality of their evidence, decisions, and executed actions at each research stage. Human-agent metrics should capture interaction quality and delegation boundaries (Lee et al., 2022; Shao et al., 2024); data-driven discovery benchmarks should use authentic tasks, expert validation, execution results, and cost (Chen et al., 2024e); closed-loop discovery should test hypothesis formation, experiment design, action, and explanatory knowledge (Jansen et al., 2024); hypothesis metrics should assess novelty, feasibility, utility, and discovery rate (Liu et al., 2025a; Si et al., 2024); and physical agents need failure modes such as atomic force microscopy (AFM) "sleepwalking" and recovery after lost task alignment (Mandal et al., 2025).

Hallucination and Faithfulness Scientific evaluation must separate useful speculation from unsupported claims. Semantic-entropy methods target confabulations where answers vary in meaning across generations (Farquhar et al., 2024); large-scale citation audits show that fabricated references already appear in scientific writing (Zhao et al., 2026); biomedical hypothesis studies require domain-specific checks for unsupported clinical or biological claims (Qi et al., 2024a); and end-to-end AI research systems raise reviewer-independence concerns when generation and evaluation are performed by related agents (Lu et al., 2026). Evaluation should therefore check evidence support, citation validity, constraint satisfaction, uncertainty calibration, and whether reviewers can assess claims independently of the agents that generated them.

Hypothesis Generation A useful hypothesis metric should ask whether an agent proposes ideas that are plausible, testable, and worth pursuing, not merely fluent (Liu et al., 2025a; Si et al., 2024). Factual alignment, hallucination checks, and feasibility tests help screen invalid hypotheses (Qi et al., 2024a; Huang et al., 2025a), while novelty and impact require comparison with prior literature and expert judgment (Wang et al., 2024a; Alkan et al., 2025; Liu et al., 2025b). However, semantic distance can mistake unfamiliar wording for novelty, while citations arrive slowly and reflect field size, visibility, and positive-result bias. Both can help prioritize ideas for review, but need a literature check with a clear time cutoff and expert assessment; neither directly measures immediate scientific impact.

Knowledge Retrieval Retrieval evaluation should move beyond asking whether an agent returns individually relevant papers. Precision, recall, F1, and ROUGE remain useful for checking document or visual-evidence relevance (Cai et al., 2024a; Qi et al., 2024b; Agarwal et al., 2024), but scientific retrieval also needs coverage of competing methods, datasets, negative evidence, and unresolved debates. Citation-network and bibliographic-coupling signals can approximate breadth (Balasubramanian, 2025), yet experts must still judge whether the retrieved evidence supports, challenges, or changes the proposed research decision.

Experimental Execution *Completion/Recovery* tracks task completion rate and failure recovery (M. Bran et al., 2024; Cornelio and Diab, 2024); AFMBench adds the physical-agent case where an

AFM agent can keep acting after losing task alignment (Mandal et al., 2025). *Outcome Accuracy* uses ROC-AUC, sensitivity/specificity, and uncertainty calibration (Monaghan et al., 2021; Yang et al., 2024c). Calibration measures how closely predicted confidence matches observed correctness, helping determine when an uncertain result needs expert review. *Procedural Accuracy* evaluates correct reasoning via shortcut resistance and hallucination-aware metrics (Li et al., 2025a; Chen et al., 2024b). For scientific agents, procedural metrics should track whether instructions were followed, tools were chosen correctly, units and calibration were handled consistently, and conclusions were supported by retrieved evidence or experimental logs. *Critical Analysis* assesses argument quality and causal traceability (Kargupta et al., 2025; Zhang et al., 2025b).

Takeaway: Benchmarks should measure whether faster agent workflows produce scientific results supported by evidence. **Faithfulness**: score citation, protocol, molecule, and claim support; **Hypotheses**: require plausible, testable, and worthwhile ideas; **Retrieval**: judge coverage and decision usefulness, not only relevance; **Execution**: track recovery, procedural fidelity, calibration, and traceability.

6 The AIScientist Workflow

6.1 Agentic Modes of Operation

Building on human–computer interaction (HCI) perspectives (Zhang et al., 2025d), we group scientific agents by autonomy: ***passive*** systems execute requested tasks (Tang et al., 2025a; Li et al., 2024f); ***collaborative*** systems embed human checkpoints (Shao et al., 2024; Swanson et al., 2025); and ***proactive*** systems initiate hypotheses, robotic actions, or end-to-end discovery (Yamada et al., 2025; Gottweis et al., 2026; Ghareeb et al., 2026; Aygün et al., 2026). These modes distinguish who initiates work, who approves consequential steps, and where scientists can intervene. A system can switch modes across stages, initiating literature search proactively while awaiting human approval before experimental execution. Multi-agent debate and MCP-style infrastructure can improve coordination (Gottweis et al., 2026; Hou et al., 2025).

6.2 HCI Across Research Stages

Hypothesis, Goal, and Planning Generation. Scientists define goals, constraints, and risk while agents retrieve literature, check novelty, refine hypotheses, and turn broad questions into executable plans (Skarlinski et al., 2025; Si et al., 2024; Baek et al., 2025). Planner and PI-style agents break goals into steps and coordinate protocols; however, scientists still approve costly experiments and clarify uncertain objectives (Schmidgall et al., 2025; Autoscience Institute, 2025; Swanson et al., 2025; Zhou et al., 2024).

Execution. Agents retrieve data, generate code, operate tools, and repair failures through retries, self-reflection, expert consultation, or human examples (Liu et al., 2024e,a; Royer, 2024; Schmidgall et al., 2025; Jansen et al., 2025). However, catching tool-selection errors, unit mistakes, and failed recovery requires editable plans, logs, checkpoints, and controlled action spaces.

Review. LLM-as-a-judge modules, peer-review simulators, and tool-based verifiers can assess novelty, rigor, reproducibility, and evidence grounding (Baek et al., 2025; Li et al., 2024f; Schmidgall et al., 2025; Lu et al., 2024a; Gottweis et al., 2026). However, self-review may repeat a model's errors; source checks and independent expert assessment remain necessary.

Resilience and Safety. Reliable human–agent collaboration requires clear authority boundaries: agents should report uncertainty, request approval for consequential actions, preserve provenance, and allow scientists to override a plan (Shao et al., 2024; Autoscience Institute, 2025; Yadav and Shahi, 2024). For proactive laboratory agents, where language or perception errors become physical commands, bounded action spaces, validators, and recovery policies are part of the interface (Mandal et al., 2025; Zhou et al., 2025b).

6.3 Failure Case Studies

AFMBench exposes "sleepwalking", where an AFM agent keeps acting after losing task alignment (Mandal et al., 2025); LIRA shows that visual ambiguity, calibration, localization, and recovery policies can derail robotic inspection (Zhou et al., 2025b); and hypothesis/reviewing agents can amplify infeasible ideas or make fabricated evidence sound coherent when the same model judges novelty and feasibility (Si et al., 2024; Lu et al., 2026). Together, these cases show why workflows need bounded actions, source and execution logs, validators, replication checks, and human approval before consequential steps.

Paper	Taxonomy	Ag.	DM.	Method	HCI.	Ben.	#Dom.
Zhang et al. (2024f)	Domain	✗	Seq.+	Train. only	✗	✓	6
Gridach et al. (2025)	Research Workflow	✓	✗	Infer. only	✓	✗	4
Luo et al. (2025b)	Research Workflow	✗	✗	✗	✗	✗	✗
Zhang et al. (2025f)	Research Workflow	✗	Seq.+	✗	✗	✗	✗
Ren et al. (2025)	Agent Composition	✓	✗	Train. & Infer.	✗	✓	6+
Wei et al. (2025)	Auto. & Domain	✓	✗	Infer. only	✗	✓	4
Hu et al. (2025b)	Data & Domain	✓	✓	✗	✗	✓	6+
Zheng et al. (2025)	Research Workflow & Auto.	✓	✗	Infer. only	✓	✓	6+
Zhou et al. (2025a)	Research Workflow	✓	✗	Infer. only	✓	✓	6+
Ours	**ML & Research Pipeline**	✓	✓	**Train. & Infer.**	✓	✓	**10**

Table 2: **Comparison of related surveys.** Ag.=Agentic AI; DM.=Data Modality; HCI.=Human-Computer Interaction; Ben.=Benchmark; #Dom.=Domains covered; Seq.+=Sequence and more; Train.=Training; Infer.=Inference; Auto.=Autonomy Level.

Takeaway: Human-AI interaction should scale oversight with consequence. Low-risk retrieval or drafting can be delegated, but hypothesis selection, execution, and review require editable plans, provenance, validators, and escalation; physical or costly actions need explicit approval and recovery policies.

7 Related Survey

This is a survey centered on representative scientific-agent research from 2024–2026.

Prior surveys examine scientific AI agents by workflow stages (Luo et al., 2025b; Zhang et al., 2025f), autonomy levels (Gridach et al., 2025; Ren et al., 2025), and domain resources (Wei et al., 2025; Hu et al., 2025b); recent EMNLP surveys cover automation-to-autonomy (Zheng et al., 2025) and AI support from hypothesis generation to publication (Zhou et al., 2025a). Table 2 compares their coverage across six dimensions. Our survey adds a **pipeline-oriented** view across modalities, methods, benchmarks, and human-AI collaboration, clarifying how multimodal scientific agents are built, where costs arise, and which human checkpoints remain necessary.

8 Limitations & Future Work

Creativity, Faithfulness, and Scientific Evidence. Brainstorming can also produce invalid molecules, unreliable protocols, fabricated citations, or unsupported biomedical claims. These failures are exposed by semantic-entropy tests, citation audits, hypothesis-support benchmarks, and independent-review settings (Farquhar et al., 2024; Zhao et al., 2026; Qi et al., 2024a; Lu et al., 2026); future systems should link claims to retrievable evidence and independent checks.

Domain-Specific Perception and Result Faithfulness. Scientific perception depends on instruments, representations, and experimental conditions. These modality-specific failures are exposed by MMSci, MicroVQA, and MaCBench through diagrams, microscopy, chemistry/materials spatial reasoning, and cross-modal synthesis (Li et al., 2024h; Burgess et al., 2025; Alampara et al., 2025); future work should pair such benchmarks with validators, uncertainty reports, and instrument logs so conclusions can be checked against measurements and their uncertainty.

Safety, Toxicity, and Laboratory Action. A mistaken suggestion can become a serious hazard when translated into laboratory action. These risks are exposed by dual-use red-teaming, broader hazard analyses, AFMBench misalignment cases, and LIRA lab-localization failures (Urbina et al., 2022; Tang et al., 2025b; Mandal et al., 2025; Zhou et al., 2025b); future systems therefore need bounded action spaces, hardware checks, permissioned tools, recovery policies, red-teaming, and human approval before high-consequence actions (Yadav and Shahi, 2024; Hou et al., 2025).

9 Conclusion

This survey connects scientific data modalities with training and inference, evaluation, and human oversight. Its practical lesson is to match autonomy to reliable feedback: preserve scientific constraints, check actions and claims independently, and involve experts where errors are costly. Progress requires reproducible scientific results alongside capable models and tools.

References

Naafey Aamer, Muhammad Nabeel Asim, Shan Munir, and Andreas Dengel. 2025. Automating ai discovery for biomedicine through knowledge graphs and llm agents. *bioRxiv*, page 2025.05.08.652829.

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2024. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10757–10765.

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024. Litllms, llms for literature review: Are we there yet? *arXiv preprint arXiv:2412.15249*.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Salma J. Ahmed and Emad A. Mohammed. 2024. Improving targeted molecule generation through language model fine-tuning via reinforcement learning. *arXiv preprint arXiv:2405.06836*.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. 2025. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5(10):952–961.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. 2025. A survey on hypothesis generation for scientific discovery in the era of large language models. *arXiv preprint arXiv:2504.05496*.

Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic Arabic medical dialogues using advanced multi-agent LLM techniques. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 11–26, Bangkok, Thailand. Association for Computational Linguistics.

AlphaProof and AlphaGeometry teams. 2024. AI achieves silver-medal standard solving International Mathematical Olympiad problems. Accessed: 2025-08-19.

Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. 2024. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Lecture Notes in Computer Science Advances in Knowledge Discovery and Data Mining*, pages 53–64. Springer.

Mehrad Ansari and Seyed Mohamad Moosavi. 2024. Agent-based learning of materials datasets from the scientific literature. *Digital Discovery*, 3(12):2607–2617.

Anthropic. 2024. Introducing the next generation of Claude.

Autoscience Institute. 2025. Meet Carl: The First AI System To Produce Academically Peer-Reviewed Research. Accessed: 2025-03-06.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston, Anton Kast, Cory Y. McLean, Peter Norgaard, Zahra Shamsi, David Smalling, James Thompson, Subhashini Venugopalan, Brian P. Williams, Chujun He, Sarah Martinson, Martyna Plomecka, Lai Wei, and 23 others. 2026. An AI system to help scientists write expert-level empirical software. *Nature*, 654(8120):909–916.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, S. Sara Mahdavi, Ellery Wulczyn, Boris Babenko, Megan Walker, Aaron Loh, Po-Hsuan Cameron Chen, Yuan Liu, Pinal Bavishi, Scott Mayer McKinney, and 17 others. 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico. Association for Computational Linguistics.

Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, and Amir Barati Farimani. 2023. Gpt-molberta: Gpt molecular features language model for molecular property prediction. *arXiv preprint arXiv:2310.03030*.

Abhinav Balasubramanian. 2025. Accelerating research with automated literature reviews: A rag-based framework. *International Journal of Multidisciplinary Research and Growth Evaluation.*, 6(2):337–342.

Hritik Bansal, Daniel Israel, Siyan Zhao, Shufan Li, Tung Nguyen, and Aditya Grover. 2024. Medmax: Mixed-modal instruction tuning for training biomedical assistants. *arXiv preprint arXiv:2412.12661*. Version 2.

Kristian G Barman, Sascha Caron, Emily Sullivan, Henk W de Regt, Roberto Ruiz de Austri, Mieke Boon, Michael Färber, Stefan Fröse, Faegheh Hasibi, Andreas Ipp, Rukshak Kapoor, Gregor Kasieczka, Daniel Kostić, Michael Krämer, Tobias Golling, Luis G. Lopez, Jesus Marco, Sydney Otten, Pawel Pawlowski, and 3 others. 2025. Large physics models: Towards a collaborative approach with large language models and foundation models. *arXiv preprint arXiv:2501.05382*.

Hritam Basak, Soumitri Chattopadhyay, Rohit Kundu, Sayan Nag, and Rammohan Mallipeddi. 2023. Ideal: Improved dense local contrastive learning for semi-supervised medical image segmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2023. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*.

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

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, and 4 others. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. *arXiv preprint arXiv:2503.13399*.

Hengxing Cai, Xiaochen Cai, Shuwen Yang, Jiankun Wang, Lin Yao, Zhifeng Gao, Junhan Chang, Sihang Li, Mingjun Xu, Changxin Wang, Hongshuai Wang, Yongge Li, Mujie Lin, Yaqi Li, Yuqi Yin, Linfeng Zhang, and Guolin Ke. 2024a. Uni-smart: Universal science multimodal analysis and research transformer. *arXiv preprint arXiv:2403.10301*.

Xibao Cai, Houtim Lai, Xing Wang, Longyue Wang, Wei Liu, Yijun Wang, Zixu Wang, Dongsheng Cao, and Xiangxiang Zeng. 2024b. Comprehensive evaluation of molecule property prediction with chatgpt. *Methods*, 222:133–141.

He Cao, Yanjun Shao, Zhiyuan Liu, Zijing Liu, Xiangru Tang, Yuan Yao, and Yu Li. 2024. Presto: Progressive pretraining enhances synthetic chemistry outcomes. *arXiv preprint arXiv:2406.13193*.

Yue Cao, Pengyi Yang, and Jean Yee Hwa Yang. 2021. A benchmark study of simulation methods for single-cell rna sequencing data. *Nature Communications*, 12(1):6911.

Haizhu Chen, Ruichong Lin, and Yu Yunfang. 2024a. Metapath chat: multimodal generative artificial intelligence chatbot for clinical pathology. *MedComm*, 5(10):e769.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024b. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

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

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025a. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PmLR.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. 2025b. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8(1):159.

Zhengyi Chen, Chunmin Zhang, Biyun Zhang, and Yifan He. 2024c. Triplet contrastive learning framework with adversarial hard-negative sample generation for multimodal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. 2024d. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *arXiv preprint arXiv:2401.12208*.

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. 2024e. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. 2024. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. 2022. Satmae: Pretraining transformers for temporal and multi-spectral satellite imagery. *arXiv preprint arXiv:2207.08051*.

Cristina Cornelio and Mohammed Diab. 2024. Recover: A neuro-symbolic framework for failure detection and recovery. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12435–12442.

Joseph Cox, Peng Liu, Skylar E Stolte, Yunchao Yang, Kang Liu, Kyle B See, Huiwen Ju, and Ruogu Fang. 2024. Brainsegfounder: towards 3d foundation models for neuroimage segmentation. *Medical Image Analysis*, 97:103301.

Miruna Cretu, Charles Harris, Ilia Igashov, Arne Schneuing, Marwin Segler, Bruno Correia, Julien Roy, Emmanuel Bengio, and Pietro Liò. 2024. Synflownet: Design of diverse and novel molecules with synthesis constraints. *arXiv preprint arXiv:2405.01155*.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480.

Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, and Xuming Hu. 2025. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. *arXiv preprint arXiv:2505.15472*.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2024. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*.

Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, Animesh Garg, and Florian Shkurti. 2024. Organa: A robotic assistant for automated chemistry experimentation and characterization. *arXiv preprint arXiv:2401.06949*.

Bernardo P. de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, Maren Lang, Uğur Şahin, Karim Beguir, and Thomas Pierrot. 2025. A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*, 7(6):928–941.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. 2025. Achieving gpt-4o level performance in astronomy with a specialized 8b-parameter large language model. *Scientific Reports*, 15(1):13751.

Marius-Constantin Dinu, Claudiu Leoveanu-Condrei, Markus Holzleitner, Werner Zellinger, and Sepp Hochreiter. 2024. Symbolicai: A framework for logic-based approaches combining generative models and solvers. *arXiv preprint arXiv:2402.00854*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151*.

Emadeldeen Eldele, Mohamed Ragab, Xu Qing, Edward, Zhenghua Chen, Min Wu, Xiaoli Li, and Jay Lee. 2025. Unifault: A fault diagnosis foundation model from bearing data. *arXiv preprint arXiv:2504.01373*.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 2022. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 15661–15676. PMLR.

Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P Brenner. 2024. Hardmath: A benchmark dataset for challenging problems in applied mathematics. *arXiv preprint arXiv:2410.09988*.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2022. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

ZhengZhao Feng, Rui Wang, TianXing Wang, Mingli Song, Sai Wu, and Shuibing He. 2024. A comprehensive survey of dynamic graph neural networks: Models, frameworks, benchmarks, experiments and challenges. *arXiv preprint arXiv:2405.00476*.

Dyke Ferber, Georg Wölflein, Isabella C Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar SM El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, and Jakob Nikolas Kather. 2024. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1):10104.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.

Christopher Fifty, Jure Leskovec, and Sebastian Thrun. 2023. In-context learning for few-shot molecular property prediction. *arXiv preprint arXiv:2310.08863*.

Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. 2025. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*.

Sin-Hang Fung, Zhenghao Zhang, Ran Wang, Chen Miao, Brian Shing-Hei Wong, Kelly Yichen Li, Chenyang Hong, Jingying Zhou, Kevin Y. Yip, Stephen Kwok-Wing Tsui, and Qin Cao. 2025. Few-shot in-context learning with large language models for antibody characterization. *bioRxiv*, page 2025.02.11.637772.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. 2023a. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36:44595–44614.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023b. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese,

Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alireza Ghafarollahi and Markus J Buehler. 2024. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 3(7):1389–1409.

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

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. 2026. A multi-agent system for automating scientific discovery. *Nature*, 655(8122):497–505.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR.

Mauro Giuffrè, Simone Kresevic, Nicola Pugliese, Kisung You, and Dennis L Shung. 2024. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. *Liver International*, 44(9):2114–2124.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, and 32 others. 2026. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496.

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. *arXiv preprint arXiv:2503.08979*.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. 2024. BLADE: Benchmarking language model agents for data-driven science. *arXiv preprint arXiv:2408.09667*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. 2024. Decompdiff: diffusion models with decomposed priors for structure-based drug design. *arXiv preprint arXiv:2403.07902*.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems 36*, 36:59662–59688.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491.

Graham Heimberg, Tony Kuo, Daryle DePianto, Tobias Heigl, Nathaniel Diamant, Omar Salem, Gabriele Scalia, Tommaso Biancalani, Shannon Turley, Jason Rock, Héctor Corrada Bravo, Josh Kaminker, Jason A. Vander Heiden, and Aviv Regev. 2023. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. *bioRxiv*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Shengxin Hong, Liang Xiao, Xin Zhang, and Jianxia Chen. 2024. Argmed-agents: Explainable clinical decision reasoning with llm disscusion via argumentation schemes. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5486–5493.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. 2025a. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22(1):82–91.

Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, Ying Chen, Chaoyang Zhang, Cheng Tan, Jie Ying, Guocheng Wu, Shujian Gao, Pengcheng Chen, Jiashi Lin, Haitao Wu, and 84 others. 2025b. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, Jingyang Yuan, Wei Ju, Luchen Liu, Tianyu Liu, Baobao Chang, and Ming Zhang. 2024a. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *arXiv preprint arXiv:2407.00468*.

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

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. 2025a. Automated hypothesis validation with agentic sequential falsifications. *arXiv preprint arXiv:2502.09858*.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, and 4 others. 2025b. Biomni: A general-purpose biomedical ai agent. *bioRxiv*.

Qian Huang, Hongyu Ren, Peng Chen, Gregor Kržmanc, Daniel Zeng, Percy Liang, and Jure Leskovec. 2023a. Prodigy: Enabling in-context learning over graphs. *Advances in Neural Information Processing Systems 36*, 36:16302–16317.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. 2023b. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6(1):74.

IBM Research. 2025. RXN for Chemistry. Accessed: 2025-07-18.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. 2024. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems 37*, 37:10088–10116.

Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. *arXiv preprint arXiv:2503.22708*.

Piyush Jha, Prithwish Jana, Pranavkrishna Suresh, Arnav Arora, and Vijay Ganesh. 2024. Rlsf: Fine-tuning llms via symbolic feedback. *arXiv preprint arXiv:2405.16661*.

Xingguang Ji, Yahui Liu, Qi Wang, Jingyuan Zhang, Yang Yue, Rui Shi, Chenxi Sun, Fuzheng Zhang, Guorui Zhou, and Kun Gai. 2025. Leanabell-prover-v2: Verifier-integrated reasoning for formal theorem proving via reinforcement learning. *arXiv preprint arXiv:2507.08649*.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Jiyue Jiang, Pengan Chen, Jiuming Wang, Dongchen He, Ziqin Wei, Liang Hong, Licheng Zong, Sheng Wang, Qinze Yu, Zixian Ma, Yanyu Chen, Yimin Fan, Xiangyu Shi, Jiawei Sun, Chuan Wu, and Yu Li. 2025. Benchmarking large language models on multiple tasks in bioinformatics nlp with prompting. *arXiv preprint arXiv:2503.04013*.

Zhihuan Jiang, Zhen Yang, Jinhao Chen, Zhengxiao Du, Weihan Wang, Bin Xu, and Jie Tang. 2024. Visscience: An extensive benchmark for evaluating k12 educational multi-modal scientific reasoning. *arXiv preprint arXiv:2409.13730*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and 15 others. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Hye-Soo Jung, Eun-Jae Lee, Dae-Il Chang, Han Jin Cho, Jun Lee, Jae-Kwan Cha, Man-Seok Park, Kyung Ho Yu, Jin-Man Jung, Seong Hwan Ahn, Dong-Eog Kim, Ju Hun Lee, Keun-Sik Hong, Sung-Il Sohn, Kyung-Pil Park, Sun U. Kwon, Jong S. Kim, Jun Young Chang, Bum Joon Kim, and 2 others. 2024. A multimodal ensemble deep learning model for functional outcome prognosis of stroke patients. *Journal of Stroke*, 26(2):312–320.

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

Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. 2025. Tree-of-debate: Multi-persona debate trees elicit critical thinking for scientific comparative analysis. *arXiv preprint arXiv:2502.14767*.

Kazumi Kasaura, Naoto Onda, Yuta Oriike, Masaya Taniguchi, Akiyoshi Sannai, and Sho Sonoda. 2025. Discovering new theorems via llms with in-context proof learning in lean. *arXiv preprint arXiv:2509.14274*.

Karl Keat, Rasika Venkatesh, Yidi Huang, Rachit Kumar, Sony Tuteja, Katrin Sangkuhl, Binglan Li, Li Gong, Michelle Whirl-Carrillo, Teri E Klein, Marylyn D. Ritchie, and Dokyoon Kim. 2024. Pgxqa:

A resource for evaluating llm performance for pharmacogenomic qa tasks. In *Biocomputing 2025*, pages 229–246. World Scientific.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, Aidan Gilson, Maxwell Singer, Amisha Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems 37*, 37:84730–84745.

Hyomin Kim, Yunhui Jang, and Sungsoo Ahn. 2025a. Mt-mol: Multi agent system with tool-based reasoning for molecular optimization. *arXiv preprint arXiv:2505.20820*.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. 2025b. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):240.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S. Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W. Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.

Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. 2025. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690.

Neelesh Kumar, Oya Aran, and Venugopal Vasudevan. 2024. Visual in-context learning for few-shot eczema segmentation. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5.

Varun Kumar, Leonard Gleyzer, Adar Kahana, Khemraj Shukla, and George Em Karniadakis. 2023. Mycrunchgpt: A llm assisted framework for scientific machine learning. *Journal of Machine Learning for Modeling and Computing*, 4(4):41–72.

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. 2023. The history and risks of reinforcement learning and human feedback. *arXiv preprint arXiv:2310.13595*.

Tian Lan, Shuquan Su, Pengyao Ping, Gyorgy Hutvagner, Tao Liu, Yi Pan, and Jinyan Li. 2024. Generating mutants of monotone affinity towards stronger protein complexes through adversarial learning. *Nature Machine Intelligence*, 6(3):315–325.

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. 2024. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Minhyeok Lee. 2023. Recent advances in generative adversarial networks for gene expression data: a comprehensive review. *Mathematics*, 11(14):3055.

Baiying Lei, Yafeng Li, Wanyi Fu, Peng Yang, Shaobin Chen, Tianfu Wang, Xiaohua Xiao, Tianye Niu, Yu Fu, Shuqiang Wang, Hongbin Han, and Jing Qin. 2024a. Alzheimer's disease diagnosis from multimodal data via feature inductive learning and dual multilevel graph neural network. *Medical Image Analysis*, 97:103213.

Mingcong Lei, Yiming Zhao, Ge Wang, Zhixin Mai, Shuguang Cui, Yatong Han, and Jinke Ren. 2025. Stma: A spatio-temporal memory agent for long-horizon embodied task planning. *arXiv preprint arXiv:2502.10177*.

Wanlu Lei, Caterina Fuster-Barceló, Gabriel Reder, Arrate Muñoz-Barrutia, and Wei Ouyang. 2024b. Bioimage. io chatbot: a community-driven ai assistant for integrative computational bioimaging. *Nature Methods*, 21(8):1368–1370.

Daniel Levine, Sacha Lévy, Syed Asad Rizvi, Nazreen Pallikkavaliyaveetil, Xingyu Chen, David Zhang, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique de Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul M. Dhodapkar, and David van Dijk. 2024. Cell2Sentence: Teaching Large Language Models the Language of Biology. *bioRxiv*.

Bin Li, Yin Li, and Kevin W Eliceiri. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems 36*, 36:28541–28564.

Guanghui Li, Shuwen Li, Cheng Liang, Qiu Xiao, and Jiawei Luo. 2024a. Drug repositioning based on residual attention network and free multiscale adversarial training. *BMC Bioinformatics*, 25(1):261.

Hao Li, Liuzhenghao Lv, He Cao, Zijing Liu, Zhiyuan Yan, Yu Wang, Yonghong Tian, Yu Li, and Li Yuan. 2025a. How to detect and defeat molecular mirage: A metric-driven benchmark for hallucination in llm-based molecular comprehension. *arXiv preprint arXiv:2504.12314*.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024b. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6071–6083.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. 2023b. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311*.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024c. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024d. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. 2024e. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*.

Michael Y Li, Emily B Fox, and Noah D Goodman. 2024f. Automated statistical model discovery with language models. *arXiv preprint arXiv:2402.17879*.

Ning Li, Liang Cheng, Lei Wang, Hui Chen, Yalu Zhang, Yunchang Yao, Jian Cheng, and Manchun Li. 2024g. Automatic labelling framework for optical remote sensing object detection samples in a wide area using deep learning. *Expert Systems with Applications*, 255:124827.

Rui Li, Andrew Wen, Jing Gao, and Hongfang Liu. 2023c. Mlgan: a meta-learning based generative adversarial network adapter for rare disease differentiation tasks. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10.

Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. 2025b. Stbench: Assessing the ability of large language models in spatio-temporal analysis. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 749–752.

Xinyu Li, Yingtong Huo, Chenxi Mao, Shiwen Shan, Yuxin Su, Dan Li, and Zibin Zheng. 2025c. Anomalygen: An automated semantic log sequence generation framework with llm for anomaly detection. *arXiv preprint arXiv:2504.12250*.

Yihe Li, Ruijie Meng, and Gregory J. Duck. 2025d. Large language model powered symbolic execution. *arXiv preprint arXiv:2505.13452*.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang Lim, and William Yang Wang. 2024h. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903*.

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

Youwei Liang, Ruiyi Zhang, Li Zhang, and Pengtao Xie. 2023b. Drugchat: towards enabling chatgpt-like capabilities on drug molecule graphs. *arXiv preprint arXiv:2309.03907*.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6555–6565.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, Boyi Zeng, Qiyuan Chen, Yuxun Miao, Bo Xue, Shu Wang, Luoyi Fu, Weinan Zhang, Junxian He, Yunqiang Zhu, and 2 others. 2023. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. 2025a. Hypobench: Towards systematic and principled benchmarking for hypothesis generation. *arXiv preprint arXiv:2504.11524*.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Jiayu Liu, Zhenya Huang, Chaokun Wang, Xunpeng Huang, Chengxiang Zhai, and Enhong Chen. 2024b. What makes in-context learning effective for mathematical reasoning: A theoretical analysis. *arXiv preprint arXiv:2412.12157*.

Junhua Liu, Fanfan Lin, Xinze Li, Kwan Hui Lim, and Shuai Zhao. 2024c. Physics-informed llm-agent for automated modulation design in power electronics systems. *arXiv preprint arXiv:2411.14214*.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024d. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021b. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2024e. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *arXiv preprint arXiv:2411.15692*.

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. 2024f. Evaluating the utilities of foundation models in single-cell data analysis. *bioRxiv*.

Xu Liu and Zhouhui Lian. 2024. Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv preprint arXiv:2412.05679*.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025b. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248*.

Yungeng Liu, Zan Chen, Yuguang Wang, and Yiqing Shen. 2025c. AutoProteinEngine: A large language model driven agent framework for multimodal AutoML in protein engineering. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 422–430, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuyang Liu, Liuzhenghao Lv, Xiancheng Zhang, Li Yuan, and Yonghong Tian. 2025d. Bioprobench: Comprehensive dataset and benchmark in biological protocol understanding and reasoning. *arXiv preprint arXiv:2505.07889*.

Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. 2020. Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22(32):18141–18148.

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

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024b. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764, Miami, Florida, USA. Association for Computational Linguistics.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. 2024c. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *arXiv preprint arXiv:2209.09513*.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024d. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. *arXiv preprint arXiv:2403.08192*.

Yong Lu, Chenxu Wang, Ze Wang, Xukun Zhang, Guifei Zhou, and Chunyan Li. 2025. Semi-supervised learning-based virtual adversarial training on graph for molecular property prediction. *Alexandria Engineering Journal*, 115:491–500.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024e. ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68, Bangkok, Thailand. Association for Computational Linguistics.

Erpai Luo, Jinmeng Jia, Yifan Xiong, Xiangyu Li, Xiaobo Guo, Baoqi Yu, Lei Wei, and Xuegong Zhang. 2025a. Benchmarking ai scientists in omics data-driven biological research. *arXiv preprint arXiv:2505.08341*.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023a. Dr.icl: Demonstration-retrieved in-context learning. *arXiv preprint arXiv:2305.14128*.

Shiyuan Luo, Juntong Ni, Shengyu Chen, Runlong Yu, Yiqun Xie, Licheng Liu, Zhenong Jin, Huaxiu Yao, and Xiaowei Jia. 2023b. Free: The foundational semantic recognition for modeling environmental ecosystems. *arXiv preprint arXiv:2311.10255*.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023c. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. 2024. Toward foundation model for multivariate wearable sensing of physiological signals. *arXiv preprint arXiv:2412.09758*.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025b. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

Rachel K. Luu and Markus J. Buehler. 2023. Bioinspiredllm: Conversational large language model for the mechanics of biological and bio-inspired materials. *arXiv preprint arXiv:2309.08788*.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2026. Prollama: A protein large language model for multi-task protein language processing. *IEEE Transactions on Artificial Intelligence*, 7(2):642–653.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. 2024a. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024b. Segment anything in medical images. *Nature Communications*, 15(1):654.

Kangyong Ma. 2025. Ai agents in chemical research: Gvim–an intelligent research assistant system. *Digital Discovery*, 4(2):355–375.

Zicheng Ma, Chuanliu Fan, Zhicong Wang, Zhenyu Chen, Xiaohan Lin, Yanheng Li, Shihao Feng, Ziqiang Cao, Jun Zhang, and Yi Qin Gao. 2025. Prottex: Structure-in-context reasoning and editing of proteins with large language models. *Journal of Chemical Information and Modeling*, 65(13):6599–6612.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems 36*, 36:46534–46594.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. 2020. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*.

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

Sepideh Maleki, Jan-Christian Huetter, Kangway V. Chuang, David Richmond, Gabriele Scalia, and Tommaso Biancalani. 2024. Efficient fine-tuning of single-cell foundation models enables zero-shot molecular perturbation prediction. *arXiv preprint arXiv:2412.13478*.

Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M. Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and N. M. Anoop Krishnan. 2025. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1):9104.

Ali Mansourian and Rachid Oucheikh. 2024. Chatgeoai: Enabling geospatial analysis for public through natural language, with large language models. *ISPRS International Journal of Geo-Information*, 13(10):348.

Nicholas Matsumoto, Hyunjun Choi, Jay Moran, Miguel E Hernandez, Mythreye Venkatesan, Xi Li, Jui-Hsuan Chang, Paul Wang, and Jason H Moore. 2025. Escargot: an ai agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning. *Bioinformatics*, 41(2):btaf031.

Andrew D. McNaughton, Gautham Krishna Sankar Ramalaxmi, Agustin Kruel, Carter R. Knutson, Rohith A. Varikoti, and Neeraj Kumar. 2024. Cactus: Chemistry agent connecting tool usage to science. *ACS Omega*, 9(46):46563–46573.

Kazem Meidani, Parshin Shojaee, Chandan K Reddy, and Amir Barati Farimani. 2023. Snip: Bridging mathematical symbolic and numeric realms with unified pre-training. *arXiv preprint arXiv:2310.02227*.

Robert Mendel, David Rauber, and Christoph Palm. 2023. Exploring the effects of contrastive learning on homogeneous medical image data. In *Informatik aktuell Bildverarbeitung für die Medizin 2023*, pages 128–133. Springer.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, and 16 others. 2024. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

Thomas F. Monaghan, Syed N. Rahman, Christina W. Agudelo, Alan J. Wein, Jason M. Lazar, Karel Everaert, and Roger R Dmochowski. 2021. Foundational statistical principles in medical research: Sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina*, 57(5):503.

Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. 2012. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594. PMID: 22716043.

Yonghyun Nam, Jaesik Kim, Sang-Hyuk Jung, Jakob Woerner, Erica H Suh, Dong-gi Lee, Manu Shivakumar, Matthew E Lee, and Dokyoon Kim. 2024. Harnessing artificial intelligence in multimodal omics data integration: paving the path for the next frontier in precision medicine. *Annual Review of Biomedical Data Science*, 7(1):225–250.

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. 2024. Aviary: training language agents on challenging scientific tasks. *arXiv preprint arXiv:2412.21154*.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. 2023. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*.

Tung Nguyen and Aditya Grover. 2024. Lico: Large language models for in-context molecular optimization. *arXiv preprint arXiv:2406.18851*.

Minhae Oh, Jeonghye Kim, Nakyung Lee, Donggeon Seo, Taeuk Kim, and Jungwoo Lee. 2025. Raise: Enhancing scientific reasoning in llms via step-by-step retrieval. *arXiv preprint arXiv:2506.08625*.

OpenAI. 2022. Introducing ChatGPT. Accessed: 2025-08-25.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, and 78 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. 2024. s^2 ip-llm: Semantic space informed prompt learning with llm for time series forecasting. In *International Conference on Machine Learning*, pages 39135–39153.

Dimitrios P Panagoulias, Maria Virvou, and George A Tsihrintzis. 2024. Evaluating llm–generated multimodal diagnosis from medical images and symptom analysis. *arXiv preprint arXiv:2402.01730*.

Himanshu Gautam Pandey, Akhil Amod, and Shivang Kumar. 2024. Advancing healthcare automation: Multi-agent system for medical necessity justification. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 39–49, Bangkok, Thailand. Association for Computational Linguistics.

Zeel B Patel, Yash Bachwana, Nitish Sharma, Sarath Guttikunda, and Nipun Batra. 2024. Vayubuddy: an llm-powered chatbot to democratize air quality insights. *arXiv preprint arXiv:2411.12760*.

Shishir Patil, Tianjun Zhang, Xin Wang, and Joseph Gonzalez. 2024. Gorilla: Large language model connected with massive apis. In *Advances in Neural Information Processing Systems 37*, pages 126544–126565.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1139 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*.

Edward O Pyzer-Knapp, Matteo Manica, Peter Staar, Lucas Morin, Patrick Ruch, Teodoro Laino, John R Smith, and Alessandro Curioni. 2025. Foundation models for materials discovery–current state and future directions. *npj Computational Materials*, 11(1):61.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024a. Large language models as biomedical hypothesis generators: A comprehensive evaluation. *arXiv preprint arXiv:2407.08940*.

Ruihua Qi, Weilong Li, and Haobo Lyu. 2024b. Generation of scientific literature surveys based on large language models (llm) and multi-agent systems (mas). In *Lecture Notes in Computer Science Natural Language Processing and Chinese Computing*, pages 169–180.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. Autoact: Automatic agent learning from scratch for qa via self-planning. *arXiv preprint arXiv:2401.05268*.

Jiacheng Qin, Dan Xu, Hao Zhang, Zhaoyu Xiong, Yejing Yuan, and Kangjian He. 2025. Btsegdiff: Brain tumor segmentation based on multimodal mri dynamically guided diffusion probability model. *Computers in Biology and Medicine*, 186:109694.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pages 9695–9717.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. 2024. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019.

Mrinal Rawat, Ambuje Gupta, Rushil Goomer, Alessandro Di Bari, Neha Gupta, and Roberto Pieraccini. 2025. Pre-act: Multi-step planning and reasoning improves acting in llm agents. *arXiv preprint arXiv:2505.09970*.

Rajat Rawat. 2024. Disasterqa: A benchmark for assessing the performance of llms in disaster response. *arXiv preprint arXiv:2410.20707*.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. 2022. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*.

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

Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. 2022. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93.

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

Jonathan Roberts, Timo Lüddecke, Rehan Sheikh, Kai Han, and Samuel Albanie. 2024. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 554–563.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571.

Yusuf H Roohani, Jian Vora, Qian Huang, Percy Liang, and Jure Leskovec. 2024. Biodiscoveryagent: An AI agent for designing genetic perturbation experiments. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Mélanie Roschewitz, Fabio De Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. 2025. Robust image representations with counterfactual contrastive learning. *Medical Image Analysis*, 105:103668.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. 2023. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*.

Loïc A Royer. 2024. Omega—harnessing the power of large language models for bioimage analysis. *Nature Methods*, 21(8):1371–1373.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, Xiaodong Shen, Ning Ye, Qiang Zhang, and Yiming Mo. 2024. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature Communications*, 15(1):10160.

Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. 2025. Llm4mat-bench: benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501.

Stuart Russell and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*, 1 edition. Prentice Hall.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. 2020. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pages 8459–8468. PMLR.

Varuni Sarwal, Viorel Munteanu, Timur Suhodolschi, Dumitru Ciorba, Eleazar Eskin, Wei Wang, and Serghei Mangul. 2023. Biollmbench: A comprehensive benchmarking of large language models in bioinformatics. *bioRxiv*, page 2023.12.19.572483.

Shogo Sawamura, Kengo Kohiyama, Takahiro Takenaka, Tatsuya Sera, Tadatoshi Inoue, and Takashi Nagai. 2025. An evaluation of the performance of openai-o1 and gpt-4o in the japanese national examination for physical therapists. *Cureus*, 17(1).

Anna C Schaar, Alejandro Tejada-Lapuerta, Giovanni Palla, Robert Gutgesell, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Mojtaba Bahrami, and Fabian J. Theis. 2024. Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv*, pages 2024–04.

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

Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, Geert Litjens, and Nadieh Khalili. 2024. Navigating the landscape of multimodal ai in medicine: a scoping review on technical challenges and clinical applications. *arXiv preprint arXiv:2411.03782*.

Douglas W Selinger, Timothy R Wall, Eleni Stylianou, Ehab M Khalil, Jedidiah Gaetz, and Oren Levy. 2024. A framework for autonomous ai-driven drug discovery. *bioRxiv*, page 2024.12.17.629024.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. 2024. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*.

Mahsa Sheikholeslami, Navid Mazrouei, Yousof Gheisari, Afshin Fasihi, Matin Irajpour, and Ali Motahharynia. 2024. Druggen: Advancing drug discovery with large language models and reinforcement learning feedback. *arXiv preprint arXiv:2411.14157*.

Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. 2024. A fine-tuning dataset and benchmark for large language models for protein understanding. *arXiv preprint arXiv:2406.05540*.

Jimeng Shi, Sizhe Zhou, Bowen Jin, Wei Hu, Runchu Tian, Shaowen Wang, Giri Narasimhan, and Jiawei Han. 2025. Hypercube-based retrieval-augmented generation for scientific question-answering. *arXiv preprint arXiv:2505.19288*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. 2024. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339, Miami, Florida, USA. Association for Computational Linguistics.

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

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Shambhavi Singh, Zejian Huang, Akshaya Kesarimangalam Srinivasan, Geordan Gutow, Bhaskar Vundurthy, and Howie Choset. 2024a. Hierarchical planning for long-horizon multi-agent collective construction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9003–9009. IEEE.

Simranjit Singh, Michael Fore, and Dimitrios Stamoulis. 2024b. Geollm-engine: A realistic environment for building geospatial copilots. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 585–594.

Michael Skarlinski, Tyler Nadolski, James Braza, Remo Storni, Mayk Caldas, Ludovico Mitchener, Michaela Hinks, Andrew White, and Sam Rodriques. 2025. FutureHouse Platform: Superintelligent AI Agents for Scientific Discovery.

Khachik Smbatyan, Tsolak Ghukasyan, Tigran Aghajanyan, Hovhannes Dabaghyan, Sergey Adamyan, Aram Bughdaryan, Vahagn Altunyan, Gagik Navasardyan, Aram Davtyan, Anush Hakobyan, Aram Gharibyan, Arman Fahradyan, Artur Hakobyan, Hasmik Mnatsakanyan, Narek Ginoyan, and Garik Petrosyan. 2025. Can ai agents design and implement drug discovery pipelines? *arXiv preprint arXiv:2504.19912*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar. 2024. Lean copilot: Large language models as copilots for theorem proving in lean. *arXiv preprint arXiv:2404.12534*.

Tao Song, Man Luo, Xiaolong Zhang, Linjiang Chen, Yan Huang, Jiaqi Cao, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, Guoqing Zhang, Fei Zhang, Weiwei Shang, Yao Fu, Jun Jiang, and Yi Luo. 2025. A multiagent-driven robotic ai chemist enabling autonomous chemical research on demand. *Journal of the American Chemical Society*, 147(15):12534–12545.

Yifan Song, Weimin Xiong, Dawei Zhu, Cheng Li, Ke Wang, Ye Tian, and Sujian Li. 2023. RestGPT: Connecting Large Language Models with Real-World Applications via RESTful APIs. *arXiv preprint arXiv:2306.06624*.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023a. Scieval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*.

Yimin Sun, Chao Wang, and Yan Peng. 2023b. Unleashing the potential of large language model: Zero-shot vqa for flood disaster scenario. In *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering*, pages 368–373.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023c. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Teo Susnjak, Peter Hwang, Napoleon Reyes, Andre LC Barczak, Timothy McIntosh, and Surangika Ranathunga. 2025. Automating research synthesis with domain-specific large language model fine-tuning. *ACM Transactions on Knowledge Discovery from Data*, 19(3):1–39.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. 2023. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91.

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

Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, Arman Cohan, Dov Greenbaum, Zhiyong Lu, and Mark Gerstein. 2025b. Risks of AI scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16(1):8317.

Xiangru Tang, Bill Qian, Rick Gao, Jiakang Chen, Xinyun Chen, and Mark B. Gerstein. 2024a. BioCoder: a benchmark for bioinformatics code generation with large language models. *Bioinformatics*, 40(Supplement_1):i266–i276.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024b. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Amitayush Thakur, George Tsoukalas, Yeming Wen, Jimmy Xin, and Swarat Chaudhuri. 2023. An in-context learning agent for formal theorem-proving. *arXiv preprint arXiv:2310.04353*.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, and Patrick T. Ellinor. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

Jie Tian, Martin Taylor Sobczak, Dhanush Patil, Jixin Hou, Lin Pang, Arunachalam Ramanathan, Libin Yang, Xianyan Chen, Yuval Golan, Xiaoming Zhai, Hongyue Sun, Kenan Song, and Xianqiao Wang. 2025. A multi-agent framework integrating large language models and generative ai for accelerated metamaterial design. *arXiv preprint arXiv:2503.19889*.

Minyang Tian, Luyu Gao, Shizhuo D. Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, and 11 others. 2024. Scicode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems*, volume 37, pages 30624–30650. Curran Associates, Inc.

Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, Naruki Yoshikawa, Samantha Corapi, Gun Deniz Akkoc, Felix Strieth-Kalthoff, Martin Seifrid, and Alán Aspuru-Guzik. 2024. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, and 14 others. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025a. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, and 11 others. 2023b. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019a. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Qinyong Wang, Zhenxiang Gao, and Rong Xu. 2023c. Exploring the in-context learning ability of large language model for biomedical concept linking. *arXiv preprint arXiv:2307.01137*.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019b. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pages 429–436, New York, NY, USA. Association for Computing Machinery.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023d. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min zhang, Qingsong Wen, Wei Ye, Shikun Zhang,

and Yue Zhang. 2024b. Autosurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems 37*, pages 115119–115145.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2023e. Instructprotein: Aligning human and protein language via knowledge instruction. *arXiv preprint arXiv:2310.03269*.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, and Zhiyong Lu. 2024c. Geneagent: Self-verification language agent for gene set knowledge discovery using domain databases. *arXiv preprint arXiv:2405.16205*.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala, and Rishita Anubhai. 2023f. Biobridge: Bridging biomedical foundation models via knowledge graphs. *arXiv preprint arXiv:2310.03320*.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, Junyi Gao, and Liantao Ma. 2025b. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 2250–2261, New York, NY, USA. Association for Computing Machinery.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025c. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35*, 35:24824–24837.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. 2023. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Xuming He, Qiang Zhang, Chenyu You, Shuangjia Zheng, Ning Ding, Wanli Ouyang, Nanqing Dong, Yu Cheng, and 3 others. 2025. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111*.

Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. 2022. Naturalprover: Grounded mathematical proof generation with language models. *Advances in Neural Information Processing Systems 35*, 35:4913–4927.

Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2023a. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. *arXiv preprint arXiv:2301.13629*.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. 2023b. Cellplm: pre-training of cell language model beyond single cells. *BioRxiv*, pages 2023–10.

Hongzhi Wen, Wenzhuo Tang, Wei Jin, Jiayuan Ding, Renming Liu, Xinnan Dai, Feng Shi, Lulu Shang, Hui Liu, and Yuying Xie. 2023c. Single cells are spatial tokens: Transformers for spatial transcriptomic data imputation. *arXiv preprint arXiv:2302.03038*.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, and 34 others. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Shican Wu, Xiao Ma, Dehui Luo, Lulu Li, Xiangcheng Shi, Xin Chang, Xiaoyun Lin, Ran Luo, Chunlei Pei, Changying Du, Zhi-Jian Zhao, and Jinlong Gong. 2025. Automated literature research and review-generation method based on large language models. *National Science Review*, 12(6):nwaf169.

Xingqi Wu, Yi Shi, Minghui Wang, and Ao Li. 2023a. Camr: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics*, 39(1):btad025.

Yilei Wu, Zijian Dong, Chongyao Chen, Wangchunshu Zhou, and Juan Helen Zhou. 2023b. Supremix: Supervised contrastive learning for medical imaging regression with mixup. *arXiv preprint arXiv:2309.16633*.

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2023c. Mathchat: Converse to tackle challenging math problems with llm agents. *arXiv preprint arXiv:2306.01337*.

You Wu and Lei Xie. 2025. Ai-driven multi-omics integration for multi-scale predictive modeling of genotype-environment-phenotype relationships. *Computational and Structural Biotechnology Journal*, 27:265–277.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Hongxin Xiang, Li Zeng, Linlin Hou, Kenli Li, Zhimin Fu, Yunguang Qiu, Ruth Nussinov, Jianying Hu, Michal Rosen-Zvi, Xiangxiang Zeng, and Feixiong Cheng. 2024. A molecular video-derived foundation model for scientific drug discovery. *Nature Communications*, 15(1):9696.

Zhichao Xiao, Yan Li, Yijie Ding, and Liang Yu. 2025. Epipdlf: a pretrained deep learning framework for predicting enhancer–promoter interactions. *Bioinformatics*, 41(5):btae716.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025*, pages 199–214. World Scientific.

Haomiao Xiong, Yunzhi Zhuge, Jiawen Zhu, Lu Zhang, and Huchuan Lu. 2025a. 3ur-llm: An end-to-end multimodal large language model for 3d scene understanding. *IEEE Transactions on Multimedia*, 27:2899–2911.

Zhitong Xiong, Fahong Zhang, Yi Wang, Yilei Shi, and Xiao Xiang Zhu. 2025b. Earthnets: Empowering artificial intelligence for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 13(4):45–78.

Derong Xu, Jingbo Zhou, Tong Xu, Yuan Xia, Ji Liu, Enhong Chen, and Dejing Dou. 2023. Multimodal biological knowledge graph completion via triple co-attention mechanism. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3928–3941. IEEE.

Leidong Xu, Danyal Mohaddes, and Yi Wang. 2024. Llm agent for fire dynamics simulations. *arXiv preprint arXiv:2412.17146*.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Carl Yang, Yang Xie, and Wenqi Shi. 2025a. Medagentgym: Training llm agents for code-based medical reasoning at scale. *arXiv preprint arXiv:2506.04405*.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. 2025b. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*.

Suryanarayan Kumar Yadav and Sanyogita Shahi. 2024. Safe human-robot collaboration in dynamic environments: An ai-powered situation awareness perspective. *Research Journal Of Medical Sciences*, 18(4):92–98.

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

Chengwei Yan, Yu Zhang, Jiuxin Feng, Heyang Hua, Zhihan Ruan, Zhen Li, Siyu Li, Chaoyang Yan, Pingjing Li, Jian Liu, and Shengquan Chen. 2025. Triple-effect correction for cell painting data with contrastive and domain-adversarial learning. *Nature Communications*, 16(1):6886.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*, pages 4006–4017.

Chih-Hsuan Yang, Benjamin Feuer, Zaki Jubery, Zi K. Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh K Singh, Arti Singh, Soumik Sarkar, Nirav Merchant, Chinmay Hegde, and Baskar Ganapathysubramanian. 2024a. Biotrove: A large curated image dataset enabling ai for biodiversity. *arXiv preprint arXiv:2406.17720*.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866.

Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter, Swarat Chaudhuri, and Dawn Song. 2024b. Formal mathematical reasoning: A new frontier in ai. *arXiv preprint arXiv:2412.16075*.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. 2023. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems 36*, 36:21573–21612.

Ruixin Yang, Dheeraj Rajagopal, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. 2024c. Confidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, Zhenpeng Man, Zhongming Liang, Zichen Wang, Yaning Li, Zheng Li, Yana Liu, Yao Tian, Wenhao Liu, Cong Li, and 32 others. 2024d. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34(12):830–845.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Hai-Cheng Yi, Zhu-Hong You, De-Shuang Huang, and Chee Keong Kwoh. 2022. Graph representation learning in bioinformatics: trends, methods and applications. *Briefings in Bioinformatics*, 23(1):bbab340.

Naruki Yoshikawa, Marta Skreta, Kourosh Darvish, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Andrew Zou Li, Yuchi Zhao, Haoping Xu, Artur Kuramshin, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. 2023. Large language models for chemistry robotics. *Autonomous Robots*, 47(8):1057–1086.

Adamo Young, Hannes Röst, and Bo Wang. 2024. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6(4):404–416.

Nicholas D Youngblut, Christopher Carpenter, Jaanak Prashar, Chiara Ricci-Tam, Rajesh Ilango, Noam Teyssier, Silvana Konermann, Patrick D Hsu, Alexander Dobin, David P Burke, Hani Goodarzi, and Yusuf H. Roohani. 2025. scbasecount: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv*, page 2025.02.27.640494.

Botao Yu, Frazier N. Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. 2024. Chemtoolagent: The impact of tools on language agents for chemistry problem solving. *arXiv preprint arXiv:2411.07228*.

Yongna Yuan, Xiaohang Pan, Xiaohong Li, Ruisheng Zhang, and Wei Su. 2025. A 3d generation framework using diffusion model and reinforcement learning to generate multi-target compounds with desired properties. *Journal of Cheminformatics*, 17(1):93.

Rasoul Zahedifar, Sayyed Ali Mirghasemi, Mahdieh Soleymani Baghshah, and Alireza Taheri. 2025. Llm-agent-controller: A universal multi-agent large language model system as a control engineer. *arXiv preprint arXiv:2505.19567*.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. 2023. Mascqa: A question answering dataset for investigating materials science knowledge of large language models. *arXiv preprint arXiv:2308.09115*.

Matineh Zavar, Hamid Reza Ghaffari, and Hamid Tabatabaee. 2025. An optimized multi-task contrastive learning framework for hifu lesion detection and segmentation. *Scientific Reports*, 15(1):29666.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. 2019. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*.

Chenhui Zhang and Sherrie Wang. 2024. Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7839–7849.

Daoan Zhang, Weitong Zhang, Yu Zhao, Jianguo Zhang, Bing He, Chenchen Qin, and Jianhua Yao. 2023. Dnagpt: A generalized pre-trained tool for versatile dna sequence analysis tasks. *arXiv preprint arXiv:2307.05628*.

Di Zhang, Jiatong Li, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. 2024a. Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B. *arXiv preprint arXiv:2406.07394*.

Gongbo Zhang, Yanting Li, Renqian Luo, Pipi Hu, Zeru Zhao, Lingbo Li, Guoqing Liu, Zun Wang, Ran Bi, Kaiyuan Gao, Liya Guo, Yu Xie, Chang Liu, Jia Zhang, Tian Xie, Robert Pinsler, Claudio Zeni, Ziheng Lu, Yingce Xia, and 6 others. 2025a. Unigenx: Unified generation of sequence and structure with autoregressive diffusion. *arXiv preprint arXiv:2503.06687*.

Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Xinyu Zhang, Fangzhi Xu, Qika Lin, Rui Mao, Erik Cambria, and Jun Liu. 2025b. Maps: A multi-agent framework based on big seven personality and socratic guidance for multimodal scientific problem solving. *arXiv preprint arXiv:2503.16905*.

Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. 2024b. Geoeval: Benchmark for evaluating llms and multi-modal models on geometry problem-solving. *arXiv preprint arXiv:2402.10104*.

Qiannan Zhang, Huafeng Wu, Li'nian Liang, Xiaojun Mei, and Jiangfeng Xian. 2024c. Robust sensor selection based on maximum correntropy criterion for ocean data reconstruction. *Frontiers in Marine Science*, 11:1467519.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024d. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Lecture Notes in Computer Science Computer Vision – ECCV 2024*, pages 169–186. Springer.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. 2024e. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, and 5 others. 2025c. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640.

Shuning Zhang, Hui Wang, and Xin Yi. 2025d. Exploring collaboration patterns and strategies in human-ai co-creation through the lens of agency: A scoping review of the top-tier hci literature. *arXiv preprint arXiv:2507.06000*.

Xiangying Zhang, Haotian Gao, Yifei Qi, Yan Li, and Renxiao Wang. 2025e. Generation of rational drug-like molecular structures through a multiple-objective reinforcement learning framework. *Molecules*, 30(1):18.

Yanbo Zhang, Sumeer A Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, Saso Dzeroski, Jesper Tegner, and Hector Zenil. 2025f. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1(1):14.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024f. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.

Mengnan Zhao, Lihe Zhang, Jingwen Ye, Huchuan Lu, Baocai Yin, and Xinchao Wang. 2024. Adversarial training: A survey. *arXiv preprint arXiv:2410.15042*.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2023. Knowledgemath: Knowledge-intensive math word problem solving in finance domains. *arXiv preprint arXiv:2311.09797*.

Zhenyue Zhao, Yihe Wang, Toby Stuart, Mathijs De Vaan, Paul Ginsparg, and Yian Yin. 2026. LLM hallucinations in the wild: Large-scale evidence from non-existent citations. *arXiv preprint arXiv:2605.07723*.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

Artem Zholus, Maksim Kuznetsov, Roman Schutski, Rim Shayakhmetov, Daniil Polykovskiy, Sarath Chandar, and Alex Zhavoronkov. 2025. Bindgpt: A scalable framework for 3d molecular design via language modeling and reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26083–26091.

Juexiao Zhou, Bin Zhang, Guowei Li, Xiuying Chen, Haoyang Li, Xiaopeng Xu, Siyuan Chen, Wenjia He, Chencheng Xu, Liwei Liu, and Xin Gao. 2024. An ai agent for fully automated multi-omic analyses. *Advanced Science*, 11(44):2407094.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023a. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, Yuka Kihara, UK Biobank Eye Vision Consortium, Naomi Allen, John E. J. Gallacher, Thomas Littlejohns, Tariq Aslam, Paul Bishop, Graeme Black, Panagiotis Sergouniotis, and 76 others. 2023b. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163.

Zekun Zhou, Xiaocheng Feng, Lei Huang, Xiachong Feng, Ziyun Song, Ruihan Chen, Liang Zhao, Weitao Ma, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, Ting Liu, and Bing Qin. 2025a. From hypothesis to publication: A comprehensive survey of AI-driven research support systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11773–11803, Suzhou, China. Association for Computational Linguistics.

Zhengxue Zhou, Satheeshkumar Veeramani, Francisco Munguia-Galeano, Hatem Fakhruldeen, and Andrew I. Cooper. 2025b. Localization, inspection, and reasoning (LIRA) module for autonomous workflows in self-driving laboratories. *Communications Chemistry*, 8(1):384.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Shaoting Zhang, and Xiaofan Zhang. 2025. MeNTi: Bridging medical calculator and LLM agent with nested

tool calling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5097–5116, Albuquerque, New Mexico. Association for Computational Linguistics.

APPENDIX

A Released Artifacts

GitHub:
https://github.com/JianJinglin/awesome-agentic-AIScientists

B Preliminary of Agentic AIScientists

B.1 Scientific Workflow and the Need for Multimodality

The scientific discovery workflow is a multifaceted and iterative process encompassing hypothesis formation, experimental design, data acquisition, computational analysis, and finally, knowledge generation. Here, we adopt the standard scientific research life cycle shown in Figure 3 to describe the task space that AIScientists need to cover.

Hypothesis Generation Hypothesis generation marks the beginning of scientific inquiry, where researchers form new research questions by identifying knowledge gaps, tracking current trends, and applying domain-specific intuition (Baek et al., 2025; Wu et al., 2025). This process relies on domain experts' ability to combine multiple complex inputs: interpreting textual content in domain-specific papers, synthesizing associated visual cues such as technical diagrams and trending charts, and formulating targeted literature queries, to spark creativity (Luo et al., 2025b; Ferrag et al., 2025). It also involves subtle, collaborative behaviors that are hard to formalize but central to scientific creativity (Agarwal et al., 2024; Eger et al., 2025). Recent work like ResearchAgent (Baek et al., 2025) shows that LLMs can help refine hypotheses through structured data integration and multi-agent feedback, simulating a brainstorming process. Their ability to understand jargon and complex context across disciplines further supports the generation of relevant and novel hypotheses, especially in interdisciplinary settings (Ferrag et al., 2025; Baek et al., 2025). Moreover, automated systems can propose unconventional research directions beyond human intuition, pushing the boundaries of discovery (Agarwal et al., 2024).

Literature Review. During the literature review, scientists synthesize existing knowledge and then contextualize new hypotheses. It involves reading various full-text articles, interpreting figures and tables, cross-referencing code or algorithm descriptions, and reconciling conflicting findings from multiple sources (Eger et al., 2025; Luo et al., 2025b). This phase often overlaps with idea creation, as new insights can reshape initial hypotheses. It serves as the analytical core of a scientist's cognitive workflow, requiring their integration of *text*, *diagrams*, *tabular data*, *figures*, and sometimes *code* examples. In this phase, researchers are often challenged by the exponential growth of publications and the need to efficiently extract pertinent information from heterogeneous sources, a task that human processes traditionally struggled to address (Eger et al., 2025; Chen et al., 2025a). Recent advancements in multimodal LLMs allow for comprehensive literature reviews by leveraging retrieval-augmented generation methods, which integrate textual discussion, graphical analysis, and code into a unified understanding of the field (Schmidgall et al., 2025). Semantic web search and citation mapping can help keep reviews current, although coverage still depends on the sources and retrieval process (Eger et al., 2025; Schmidgall et al., 2025).

Method Design. The method design phase requires scientists to convert theoretical ideas into actionable workflows by planning, developing, or adapting computational models, algorithms, and theoretical frameworks to address the identified research questions (Gridach et al., 2025). This stage involves multiple modalities: *symbolic* representation in the mathematical equations, *visual* info via block diagrams or architectural sketches, and *code* that translates concepts into working models (Eger et al., 2025; Gridach et al., 2025). Multimodal LLMs are particularly well-suited to this challenge as they can take and generate conceptual diagrams, write code, and produce formatted equations (Gridach et al., 2025). Recent studies have demonstrated that when these models are coupled with iterative human feedback, the method design not only becomes more robust but also allows for rapid reconfiguration when initial hypotheses need modification (Ren et al., 2025; Baek et al., 2025).

Executing and Analyzing Experiments. The experiments phase involves implementing methods to validate hypotheses and then collecting data and conducting analysis, including the actual implementation, simulation, or laboratory testing of the designed methods (Gridach et al., 2025; Eger et al., 2025). Researchers then gather information from a range of sources to assess the performance of the methods, including laboratory instrument

data (e.g., microscopy images (Panagoulias et al., 2024), time series (Pan et al., 2024), sensor logs (Li et al., 2025c), and 3D structures (Xiong et al., 2025a)), public repositories (e.g., multiomics data (Nam et al., 2024), genomic sequences (Liu et al., 2024a), optical remote sensing (Li et al., 2024g)), and simulation platforms (Xu et al., 2024), requiring scientists to handle diverse formats. To support this, recent work in multimodal LLMs emphasizes the use of these models to automate not only the ingestion of data from these heterogeneous sources but also the preprocessing steps, including noise filtering, feature extraction, normalization, and transformation from one format to another (Luo et al., 2025b; Chen et al., 2025a; Schmidgall et al., 2025). Moreover, these models can be employed to generate detailed analyses and visualizations that help interpret complex experimental results (e.g., *numerical tables*, *visual plots*, and *textual logs*), as well as to identify key patterns and anomalies across multimodal datasets (Gridach et al., 2025; Eger et al., 2025). Beyond software, automated systems also facilitate the design, execution, and refinement of experiments by orchestrating diverse components such as scientific software, physical hardware (e.g., pipettes), and real-time data acquisition systems (Boiko et al., 2023).

Writing Papers and Reports. The final stage of the scientific workflow involves the synthesis of research findings into cohesive papers and technical reports, a process that merges narrative text with figures, tables, and code excerpts (Schmidgall et al., 2025; Chen et al., 2025a). In the writing phase, the challenge is to translate complex experimental outcomes and analyses into clear, accessible, and reproducible documentation that both communicates insights and allows for independent verification (Wilkinson et al., 2016; Lu et al., 2024a). Recent agentic LLM frameworks have demonstrated significant advancements in automatically generating parts of the manuscript, such as drafting different sections, curating related works, and providing reference recommendations based on real-time web queries (Narayanan et al., 2024; Schmidgall et al., 2025). For example, systems like Agent Laboratory have shown that by integrating automated literature retrieval modules, LLMs can format citations and generate composite figures that include interactive visualizations and embedded code snippets, streamlining the entire writing process (Schmidgall et al., 2025; Baek et al., 2025).

The traditional research cycle typically requires researchers to manually shuttle between multiple tools for hypothesis generation, literature review, method design, experimentation, data collection, and report writing, often introducing inefficiencies and potential errors, especially as the complexity of tasks and the volume of multimodal information grow (Baek et al., 2025; Eger et al., 2025). However, this is where *Agentic Multimodal Large Language Models (MLLMs)* emerge as transformative tools. This integration can improve research productivity by automating repetitive tasks and helping researchers focus on conceptual challenges (Eger et al., 2025). Nevertheless, scientific rigor still depends on independent validation, reproducible execution, and expert review of these outputs (Gridach et al., 2025). Their value should therefore be assessed through verified outcomes rather than output volume alone.

B.2 Multimodal Agent Composition

As discussed in the previous section, the workflow of **an AIScientist** involves diverse tasks requiring interaction with multimodal data, tools, and representations. To effectively support such tasks, **a scientist agent** must be capable of sensing, interpreting, and acting on complex multimodal data in both digital and physical environments. Formally, anything that perceives the environment through sensors and acts on that environment through actuators can be considered **an agent** (Russell and Norvig, 1995). A **multimodal agent** is designed to integrate and process heterogeneous data, such as text, images, audio, and video, in a seamless and context-aware manner (Durante et al., 2024). These agents are built on pre-trained foundation models such as large language models (LLMs) and vision-language models (VLMs), which provide semantic understanding and generative capabilities (Durante et al., 2024). To enable flexible and adaptive behavior, multimodal agents are designed with **a modular architecture**, where each module handles a specific function: **(1) Environment & Perception**, **(2) Agent Learning**, **(3) Memory**, **(4) Action**, and **(5) Cognition**. These modules collectively support the core capabilities of an AIScientist agent. Below, we detail these modules to illustrate the AIScientist agent's internal composition.

Environment & Perception This module enables the agent to sense and interpret multimodal data from its environment, including scientific liter-

ature, instrument readouts, visual scenes, and user queries. Modern perception models are not only tasked with processing low-level signals but also with incorporating task-planning and skill observation functionalities to facilitate a semantically interpreted environment (Durante et al., 2024). To achieve efficient perception, the module leverages state-of-the-art VLMs such as CLIP, which can transform visual data into language features. For example, images captured from laboratory experiments or field environments are encoded, and salient features are extracted that highlight objects, spatial relationships, and contextual cues. These features are then combined with linguistic inputs provided by the user or retrieved from a task plan, allowing the agent to ground its interpretation of the scene on both perceptual data and prior knowledge (Durante et al., 2024). Moreover, the agent uses multimodal inputs to determine which skills, tools, or laboratory instruments may be most relevant for the current task.

Learning. The Agent Learning module is designed to improve and refine the agent's capabilities over time through an interactive learning paradigm. Modern interactive learning strategies usually wrap a pre-trained large model (e.g., LLMs or VLMs) in the agent, which is designed to actively collect user feedback (Durante et al., 2024). Agent learning proceeds in three stages: retrieval, generation, and light training. First, the agent retrieves relevant evidence from its multimodal knowledge base to prepare for the following content generation. Next, it combines this retrieved knowledge with user feedback to construct prompts that guide the pretrained model in generating responses or actions, often involving multiple cycles of self-verification and rewriting. Finally, the agent compares its generated outputs against expert evaluations and uses methods such as self-supervised learning, reinforcement learning, or imitation learning to refine its own behavior (Durante et al., 2024). In this way, if a scientific experiment yields unexpected results, the agent registers the discrepancy along with the context and updates its internal policy.

Memory. The memory module provides statefulness and continuity throughout the agent's operation, and it plays an important role in both short-term context management and long-term knowledge retention (Durante et al., 2024). Memory is implemented by leveraging a combination of episodic memory and semantic memory. Episodic memory captures transient interactions and events, allowing the agent to recall recent experiences that are relevant to the task at hand. This is especially useful in iterative experimental settings where feedback loops are critical for refining next actions (Ren et al., 2025). On the other hand, semantic memory stores generalized facts and domain-specific knowledge that the agent has learned over time. Such knowledge is typically incorporated during the pre-training phase of LLMs and can be periodically updated in light of new evidence. Thus, the memory module is essential for long-term task planning, which allows the agent to monitor its performance over periods and adjust its strategies by referencing previous successes and failures. Through an integrated memory retrieval mechanism, the agent can effectively operate over a sequence of multi-turn interactions, maintaining coherence and building upon prior context in ways that are critical for complex decision-making and adaptive behavior (Durante et al., 2024).

Action. The Agent Action module serves as the execution engine, translating the outputs of the previous modules into concrete actions that affect either digital systems or physical environments (Zahedifar et al., 2025). This module is responsible for converting high-level plans, learned policies, and contextual cues into a set of executable commands that drive external tools, APIs, protocols, or physical actuators (Durante et al., 2024). In **digital domains**, the action module may interface with software tools such as WebSearch APIs (Yu et al., 2024) that provide up-to-date scientific data, PythonREPL environments (Ruan et al., 2024) for interactive code execution, domain-specific tools (Edwards et al., 2022; IBM Research, 2025) for molecular structure prediction and chemical reaction simulation. For example, systems like PaperRobot (Wang et al., 2019a), ChemCrow (M. Bran et al., 2024), AIScientist(Lu et al., 2024a), and Biomni (Huang et al., 2025b) demonstrate how LLMs can operate software tools, forming virtual AIScientists. In **physical contexts**, the action module extends to the coordination of robotic platforms and other hardware systems, enabling the execution of real-world experiments such as solution preparation, pH measurement, or even more complex organic reactions (e.g., CMU Co-scientist (Boiko et al., 2023), CLAIRIFY (Yoshikawa et al., 2023), ChemAgents (Song et al., 2025), and ORGANA (Darvish et al., 2024)). Embodied applications

leverage task-specific controllers that transform the high-level directives of the agent into low-level motor commands (Durante et al., 2024). A significant aspect of the action module is its integration with the **environmental feedback loop**. Actions are continuously monitored and adjusted based on real-time inputs and error signals. This closed-loop design ensures that the system can recover from failures, dynamically update its action plans, and maintain reliability despite uncertainties in complex tasks (Durante et al., 2024).

Cognition. The Cognition module is the "brain" of the agent, underpinning complex reasoning, decision-making, and planning processes. It orchestrates input from Environment and Perception, Agent Learning, and Memory to generate high-level strategies (Durante et al., 2024). This module includes chain-of-thought reasoning (Wei et al., 2022), hierarchical planning (Singh et al., 2024a; Ren et al., 2025), and self-correction (Madaan et al., 2023). Typically, the cognition module is driven by LLMs that decompose complex instructions into sub-tasks and reason about dependencies between consecutive steps.

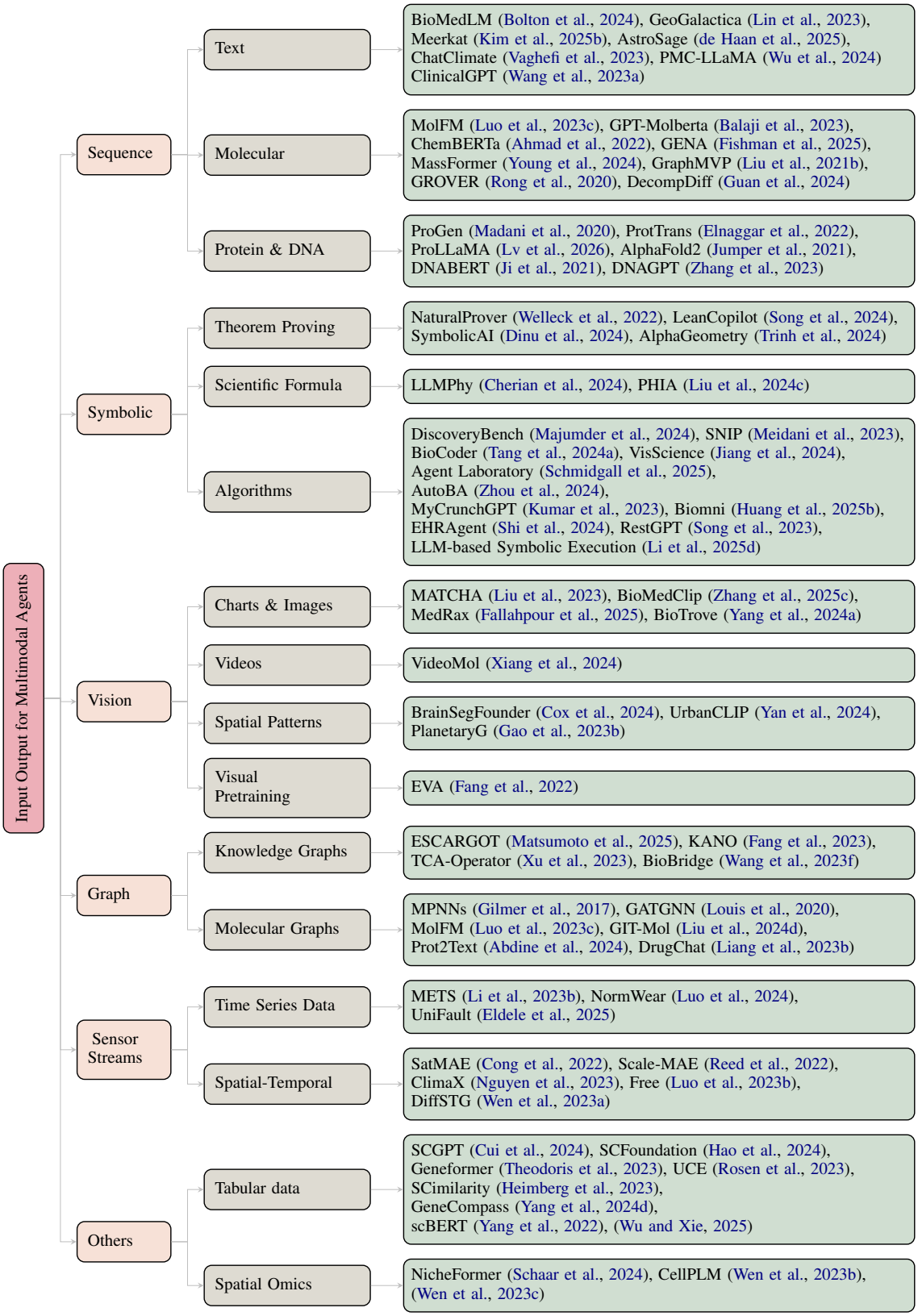


Figure 4: Taxonomy of Data Modality for Scientific MLLM Agents.

C More Details on I/O for Scientific MLLMs

C.1 Taxonomy and Overview

We categorize the input and output data modalities commonly used in scientific multimodal large language models, shown in Figure 4.

C.2 More Details for Knowledge Fusion Strategies

Models like GPT-4o show that a single model can handle multiple modalities, including text, images, and other data sources. This ability relies on multimodal knowledge fusion, which enables one model to synthesize and generate diverse data forms within a single unified architecture. This capability is important for scientific reasoning. For example, explaining a circuit or cell diagram requires both the image and the caption. Missing one can lead to incorrect or incomplete answers. Recent studies have explored diverse forms of cross-modality fusion in scientific domains, including sequence and image integration (Smbatyan et al., 2025; Ghafarollahi and Buehler, 2025; Burgess et al., 2025; Tu et al., 2024; Zhang et al., 2025c; Ma et al., 2024b; Lei et al., 2024b; Chen et al., 2024d; Royer, 2024; Lu et al., 2024c; Li et al., 2023a), sequence–graph reasoning (Selinger et al., 2024; Xu et al., 2023; Fang et al., 2023), sequence and tabular data (de Almeida et al., 2025; Levine et al., 2024), Spatial omics (Schaar et al., 2024; Wen et al., 2023b,c).

Multimodal knowledge fusion can generally be understood through three main paradigms, **Early**, **Intermediate**, and **Late Fusion**, as summarized in Table 3. The most popular strategy for multimodal knowledge fusion is **Intermediate Fusion**, where each modality is first passed through its own encoder, and the resulting feature vectors are merged before being fed into the final prediction layers (Schouten et al., 2024). Simple concatenation of these embeddings is a common practice. However, it often fails to capture deeper cross-modal interactions. To address this, more effective strategies incorporate attention mechanisms to weight complementary signals (Xu et al., 2023), or employ latent-space alignment to enforce coherent joint representations (Zhang et al., 2025c). In contrast, **Late Fusion** serves as the second common choice. This strategy trains each modality separately and then merges the models' outputs to make the final decision. Currently, it either takes a weighted average of the model predictions or feeds them into a traditional learner such as a random forest or a regression model (Krones et al., 2025; Jung et al., 2024). The least common strategy is **Early Fusion**, which merges raw inputs before any feature encoding occurs (Krones et al., 2025; Schouten et al., 2024). This requires that heterogeneous inputs share a common spatial or semantic alignment (e.g., clinical variables marked on image pixels), allowing joint feature extractors to process them simultaneously. Although simple concatenation and graph-based (Lei et al., 2024a; Fang et al., 2023) or cross-modality alignment networks (Wu et al., 2023a) have been proposed, the requirement for precise registration across modalities remains a significant challenge (Krones et al., 2025; Schouten et al., 2024). Future advancements must focus on scalable architectures that can capture complex interdependencies across modalities and adapt to domain-specific requirements.

Strategy	Fusion Timing	Operation	Interaction	Challenge	Cost Profile	Examples
Early Fusion	Before encoding	Raw-input alignment	Deepest	Strict registration requirement	Joint preprocessing and encoder training can be costly	DMGNN (Lei et al., 2024a), KANO (Fang et al., 2023)
Intermediate Fusion	Between encoder and prediction layer	Concatenation, attention, latent-space alignment	Medium to high	Designing efficient fusion modules	Medium to high; modality encoders plus cross-modal layers add GPU memory and tuning cost	TCA-Operator (Xu et al., 2023), BiomedCLIP (Zhang et al., 2025c), CAMR (Wu et al., 2023a)
Late Fusion	After model output	Prediction averaging, traditional ML	Weakest	Lacks deep inter-modality interaction	Modular, but running separate encoders can add cost	(Jung et al., 2024)

Table 3: Comparison of Multimodal Knowledge Fusion Strategies, including practical computational cost profiles.

D More Details on Methods

D.1 Taxonomy and Overview

Figure 5 provides a taxonomic overview of the methodological landscape. Tables 4 and 5 complement this by contrasting each strategy based on its key mechanism, scientific application, advantages, limitations, and practical cost profile across training methods (SFT, RL, CL/AL) and inference techniques (RAG, Planning, Self-Correction), serving as a reference guide for designing scientific MLLM agents.

D.2 Detailed Agent Training Methodologies

D.2.1 Supervised Fine-Tuning

Adapting and fine-tuning a foundation MLLM often leads to superior results compared to training a model from scratch, with costs that depend on model size, data, hardware, and tuning strategy (Maleki et al., 2024; Wei et al., 2021). Thus, supervised fine-tuning (SFT) has become the most prevalent strategy for customizing foundation models to specialized tasks (Xu et al., 2025a). Here, we introduce common SFT methods, including full supervised fine-tuning and parameter-efficient fine-tuning (PEFT).

SFT begins with selecting an appropriate foundation model and curating a task-relevant dataset of input–output pairs, such as biomedical image–caption annotations (Li et al., 2023a; Bansal et al., 2024). In full SFT, all model weights are updated, and training proceeds until evaluation metrics indicate satisfactory alignment with the task, after which the model can be frozen (Belouadi et al., 2023; Luu and Buehler, 2023). Many pretrained models have been fine-tuned to solve molecular attribute prediction tasks (de Almeida et al., 2025), protein design tasks (Lv et al., 2026), and therapeutic downstream tasks (Wang et al., 2025a). Additionally, instruction fine-tuning, a variant of full SFT, leverages natural language instructions in conjunction with input–output pairs to help models better understand the inputs' intent (Wang et al., 2025a; Bansal et al., 2024; Lu et al., 2024c).

Parameter-Efficient Fine-Tuning (PEFT) offers an alternative that significantly reduces computational and storage costs. In Low-Rank Adaptation (LoRA) fine-tuning, only a small set of low-rank adapter parameters is trained, while the base model weights remain frozen (Ma, 2025). For example, ScDCA (Maleki et al., 2024) incorporated a drug-conditional adapter into scGPT (a pretrained single-cell sequence foundation model), and MMed-RAG (Xia et al., 2024) applied LoRA adapters to align

Strategy	Mechanism	Advantages	Limitations	Scientific Examples	Cost Profile
Supervised Fine-Tuning (SFT)	Updates model weights using input-output data pairs.	More economical than training from scratch and dominant for task-specific adaptation; Simpler implementation.	Highly data-dependent & yields limited gains in factual correctness.	Molecular property prediction, Protein design, Biomedical VQA.	Low to medium for PEFT; medium to high for full SFT, depending on backbone size and expert labels.
Reinforcement Learning (RL)	Optimizes sequential decision by maximizing rewards from an environment.	Provides structured feedback (human, simulation, trajectory).	High computational cost and probably long runtimes; expensive on RLHF.	Preference alignment; molecular optimization; formal proof generation.	Often high; rollouts, verifiers, simulators, and expert feedback can dominate the budget.
Contrastive Learning (CL)	Self-supervised learning that pulls positive pairs closer and pushes negative pairs apart in the embedding space.	Uses large amounts of unlabeled data; enhances representation quality and generalizability.	Performance heavily relies on the quality of data augmentation and positive & negative pair selection.	Cross-modal representation learning (e.g., protein-molecule pairs, image-text); Data-scarce domains.	Medium to high; large batches, paired data pipelines, and GPU memory are common bottlenecks.
Adversarial Learning (AL)	Uses adversarial perturbations or generator–discriminator objectives.	Can improve robustness when perturbations remain scientifically meaningful.	Can slow down training; Risks unstable training and failure to converge.	Distinguishing chemically similar molecules.	Medium to high; dual-model training and instability can require repeated runs.

Table 4: Comparison of Scientific MLLM Agent Training Strategies. Cost profiles are qualitative because GPU-hour and expert-label budgets are rarely reported consistently.

Strategy	Mechanism	Advantages	Limitations	Scientific Examples	Cost Profile
Integrating Domain Knowledge	Grounds MLLM output using external knowledge sources (RAG, KG, ICL) for factual correctness.	Improves factual accuracy and reduces hallucination; Provides up-to-date context.	RAG/KG performance limited by source quality & coverage; ICL is sensitive to example format and incurs long-context cost.	**RAG:** Clinical guideline adherence; **KG:** Research proposal planning; **ICL:** Few-shot learning for symbolic proofs.	Low to medium per query; indexing, KG construction, and longer contexts can dominate system cost.
Complex Workflow Handling	Schedules agents' actions and reasoning into logical, multi-step processes.	Enables execution of complex, long-chain scientific tasks (e.g., design, query, synthesis).	Careful orchestration design and tool selection; sensitive to error accumulation; Increased cost.	**Planning:** Designing multi-step experiments. **Tool Learning:** Interfacing with simulation software or lab instruments.	High latency; tool calls, simulations, and robotic time grow with plan depth.
Improving Accuracy	Enables agents to evaluate their own outputs and maintain context.	Enhances reasoning reliability and continuity; Promotes transparency and trustworthiness.	Increases computational load due to iterative loops; Memory retrieval can become noisy.	**Self-Critic:** Clinical explainability & hallucination reduction. **External Memory:** Storing scientific outputs.	Medium to high; repeated calls, critique loops, and memory retrieval often multiply inference cost.
Multi-Agent System	Decomposes tasks and coordinates specialized agents.	Leverages specialized expertise and memory; Improves output accuracy via debate and voting.	Introduces new coordination challenges like group bias; Increases computational cost.	Collaborative research proposal generation; Multi-round discussion for consensus diagnosis.	Potentially high; depends on agents, rounds, tools, and communication overhead.

Table 5: Comparison of Scientific MLLM Agent Inference Strategies, including qualitative cost profiles.

MLLMs with human preferences in medical visual question answering and report generation. Beyond LoRA, methods like NEFTune introduce controlled noise into embedding vectors to enhance generalization, and have shown promising results in tasks such as automated research synthesis, though their use in scientific agents remains rare (Susnjak et al., 2025).

Despite their versatility, several studies note that SFT alone often yields limited improvements in factual accuracy (Belouadi et al., 2023; Maleki et al., 2024). Greater gains in this dimension are typically achieved through strategies that incorporate explicit domain knowledge, such as retrieval-augmented generation or knowledge graph integration, as we will discuss in subsection 4.2.1.

D.2.2 Reinforcement Learning with Feedback

Foundation models show promise when trained with reinforcement learning (RL) to solve some complex questions (AlphaProof and AlphaGeometry teams, 2024). RL refers to modeling learning to make sequential decisions to maximize rewards through interaction with the environment (Ahmed and Mohammed, 2024; Sheikholeslami et al., 2024). RL requires different rewards, such as expert preferences, simulations, synthetic rewards, and agent evaluations.

Preference Feedback Reinforcement learning from human feedback (RLHF) aligns models with preference judgments, including those used in ChatGPT (Ouyang et al., 2022; OpenAI, 2022; Lambert et al., 2023; Sun et al., 2023c; Giuffrè et al., 2024). Direct preference optimization (DPO) is a related preference-training objective (Rafailov et al., 2023). ResearchAgent generates and revises ideas using academic graphs and collaborative reviewing agents instantiated from human-aligned LLMs; this is inference-time feedback, not a newly reported RLHF training procedure (Baek et al., 2025). MAVIS uses an automatic data engine for visual mathematics instruction and reasoning data, followed by DPO to improve chain-of-thought reasoning (Zhang et al., 2024e; Wei et al., 2022). Its reasoning paths should not be described as human-annotated. However, expert preference collection can be expensive, and preference alignment alone does not establish scientific validity.

Simulation Feedback & Verified Feedback This approach provides a way to optimize MLLMs by directly incorporating domain-specific calculations or physics-based simulators into the training loop (Zhou et al., 2023a; Jha et al., 2024). Several agentic systems have been developed that use multi-turn reasoning or simulation environments for scientific tasks. For example, (Jha et al., 2024) and Leanabell-Prover-V2 (Ji et al., 2025) used symbolic feedback to guide LLMs through formal verification steps, while AgentHospital simulates medical environments for agent training (Li et al., 2024c). AtomAgents (Ghafarollahi and Buehler, 2025) and (Tian et al., 2025) apply physics-based simulation and theoretical computation for materials property prediction, integrating simulation results directly into decision-making. In molecular sciences, simulation feedback enables property-aware generation. BindGPT (Zholus et al., 2025), DrugGen (Sheikholeslami et al., 2024), and other works (Yuan et al., 2025; Ahmed and Mohammed, 2024) generate drug-like molecules, simulate binding or stability scores, and use them as reward signals for further refinement. Simulation feedback provides structured signals but depends on the validity of the simulator or verifier. It can also require costly, long-running computations, particularly in physical simulations (Zhang et al., 2025e; Sheikholeslami et al., 2024).

Trajectory Learning Trajectory optimization enhances feedback by evaluating intermediate steps, not just final outputs, and is especially effective for improving MLLM reasoning in science tasks (Xu et al., 2025a). By storing and reusing successful trajectories, models can learn efficient multi-step strategies for complex workflows such as multi-molecule synthesis or sequential materials experiments (Cretu et al., 2024). However, this method is underexplored in agentic AIScientists, and the construction cost is high.

D.2.3 Contrastive & Adversarial Learning

In scientific domains, labeled datasets are expensive to obtain (Roschewitz et al., 2025; Xiao et al., 2025; Huang et al., 2023b). Contrastive learning can learn representations from unlabeled data (Zhou et al., 2023b; Azizi et al., 2023; Chen et al., 2020). The approach forms positive pairs and selects negative examples, reducing the need for manual labels. Examples include protein–molecule pairs (Gao et al., 2023a), and enhancer–promoter interactions in DNA (Xiao et al., 2025).

Contrastive learning (CL) maps examples into a shared representation space, bringing related pairs closer and separating unrelated pairs. Similarity scores and training losses depend on the task (Xiao et al., 2025; Yan et al., 2025; Li et al., 2021). SimCLR builds positive image pairs using photometric and geometric transformations (Zavar et al., 2025; Chen et al., 2020). Counterfactual pairs (Roschewitz et al., 2025), hard negatives (Chen et al., 2024c; Wu et al., 2023b), multi-task objectives (Zavar et al., 2025; Mendel et al., 2023), and local contrastive objectives (Mendel et al., 2023; Basak et al., 2023) offer other ways to shape representations. DINO is a related self-supervised approach based on self-distillation (Oquab et al., 2023). However, in scientific tasks, these choices must preserve relevant structure; arbitrary augmentations or poorly chosen negatives can erase meaningful differences (Xiao et al., 2025).

Adversarial training introduces challenging perturbations to improve robustness (Lan et al., 2024; Zhao et al., 2024), while generative adversarial networks (GANs) use a distinct generator–discriminator objective (Yan et al., 2025; Lee, 2023; Salimans et al., 2016). A GAN trains a generator to produce samples and a discriminator to distinguish generated from real data (Li et al., 2024a, 2023c). Robustness is particularly important when distinguishing chemically similar molecules with different bioactivity (Lu et al., 2025) or predicting proteins' binding-affinity changes upon mutation (Lan et al., 2024). However, adversarial training can increase computational cost (Xiao et al., 2025) and may even fail to converge, making the overall training more difficult and less stable (Salimans et al., 2016).

D.3 Detailed Agent Inference Methodologies

D.3.1 Integrating Domain Knowledge

Incorporating domain knowledge into agentic MLLMs is critical for improving output reliability and reducing hallucinations. Retrieval-Augmented Generation (RAG), graphs, and in-context learning (ICL) provide effective means to align agents with scientific knowledge sources.

RAG RAG-based methods (Ghafarollahi and Buehler, 2025) improve factual accuracy by retrieving domain-relevant evidence. In biomedical settings, RAG on guideline documents im-

proves the alignment with clinical standards (Wang et al., 2025c,b; Lu et al., 2024b), while AutoProteinEngine (Liu et al., 2025c) demonstrates the use of literature retrieval to integrate protein engineering knowledge. Other systems employ real-time search engines to enrich responses with up-to-date information (Lu et al., 2024e; Ghafarollahi and Buehler, 2024; Lei et al., 2024b). Beyond direct retrieval, task-specific strategies have been explored, such as i-MedRAG (Xiong et al., 2024), LLM Instructor (Liu et al., 2024e), and Hypercube-Based RAG (Shi et al., 2025). However, RAG methods remain limited by the coverage and quality of the retrieval corpus, which can result in the propagation of irrelevant or misleading information.

Knowledge Graph (KG) Knowledge graphs provide additional constraints for agent reasoning. For example, (Selinger et al., 2024) and (Aamer et al., 2025) combine LLMs with biomedical graphs to autonomously design research proposals and plan experiments. ESCARGOT (Matsumoto et al., 2025) extends this by enabling code execution to dynamically query scientific knowledge graphs, allowing computational experiments within the reasoning loop. However, the scalability of knowledge-graph methods is constrained by graph size, and maintaining accurate and up-to-date knowledge graphs is quite challenging.

In-Context Learning (ICL) ICL (Brown et al., 2020) is a simple and powerful strategy, where MLLMs can learn from a few demonstrations of question-answer pairs within the context and then generate answers to a new query (Thakur et al., 2023; Dong et al., 2022; Wei et al., 2023; Huang et al., 2023a), all without updating model parameters. This approach has proven effective not only for scientific sequence data (Wang et al., 2023c) but also for multimodal data, including symbolic theorem proving (Kasaura et al., 2025; Thakur et al., 2023; Cherian et al., 2024), geospatial analysis (Singh et al., 2024b; Mansourian and Oucheikh, 2024), medical image segmentation (Kumar et al., 2024), and diverse protein and molecular tasks (Ma et al., 2025; Li et al., 2025d; Fifty et al., 2023; Cai et al., 2024b; Li et al., 2024b; Nguyen and Grover, 2024).

Demonstration pairs play a central role in influencing ICL performance (Dong et al., 2022), with outcomes further shaped by how demonstrations are selected, formatted, and ordered. Specifically, the format, order, diversity of demonstrations, and their similarity to the query (Fung et al., 2025; Liu et al., 2021a) substantially affect response quality. Using demonstrations that closely resemble the query has become a widely adopted method in ICL (Liu et al., 2024b; Luo et al., 2023a; Liu et al., 2021a), generating strong results across domains such as biological sequence classification (Fung et al., 2025), pathology image classification (Ferber et al., 2024), molecular notation (Li et al., 2025d), molecular optimization (Nguyen and Grover, 2024), and molecular property prediction (Cai et al., 2024b; Guo et al., 2023; Fifty et al., 2023).

In practice, researchers tend to increase the number of demonstrations to improve performance (Fung et al., 2025; Wang et al., 2023e). However, excessive demonstrations can also harm performance (Dong et al., 2022; Bertsch et al., 2024), a challenge that can be partially addressed through long-context solutions. Thus, the trade-off between performance gains and computational costs remains a key consideration (Hu et al., 2025a).

D.3.2 Complex Workflow Handling

Scientific tasks often involve long-chain reasoning processes, such as designing an experiment, querying databases, and synthesizing results, where agents need planning and tool use to support multi-step reasoning (Oh et al., 2025; Shi et al., 2025).

Planning Frameworks such as ReAct (Yao et al., 2022), Pre-Act (Rawat et al., 2025), and AutoAct (Qiao et al., 2024) combine reasoning with actions, enabling hierarchical workflows where models plan before executing tool calls (Youngblut et al., 2025). RAISE decomposes problems into logical queries and retrieval steps (Oh et al., 2025), while LLM-RDF integrates chain-of-thought (Wei et al., 2022) with domain tools (Bayesian optimizers, interpreters, robotic arms) for closed-loop chemical synthesis (Ruan et al., 2024). Other strategies, such as Monte Carlo Tree Search for guided refinement (Zhang et al., 2024a) and LLM planners for biomedical workflows (Liu et al., 2024e), further strengthen structured planning but may face the challenge of high costs.

Tool Learning Agents increasingly rely on domain-specific tools for scientific discovery (Fallahpour et al., 2025; McNaughton et al., 2024; Huang et al., 2024b). Early systems such as ChemCrow (M. Bran et al., 2024), BioDiscoveryAgent (Roohani et al., 2024), and Omega (Royer, 2024)

connect external APIs (e.g., LangChain, PubMed, Visual tools) to provide real-time access to literature, databases, and visual understanding. Biomni (Huang et al., 2025b) and AutoBA (Zhou et al., 2024) show large-scale tool integration and adaptive selection of biomedical tools. Advanced mechanisms such as MeNTi's nested calls (Zhu et al., 2025) highlight the need for compositional and flexible tool orchestration. Technical documentation is a valuable resource for using laboratory hardware (Boiko et al., 2023).

Multi-step Reasoning Chain-of-thought improves step-wise inference in scientific tasks (Chen et al., 2025b; Huang et al., 2024b). Role-specialized agents with domain tools yield grounded results, for example, MT-Mol uses RDKit for chemical reasoning to improve molecular synthesis design (Kim et al., 2025a). Also, hierarchical workflows, such as MedAgent-Pro, make medical reasoning more explainable (Wang et al., 2025c). Iterative chains of verification enable cross-checking and predicting material attributes from literature (Ansari and Moosavi, 2024). However, these methods remain sensitive to early errors and limited by models' hallucinations.

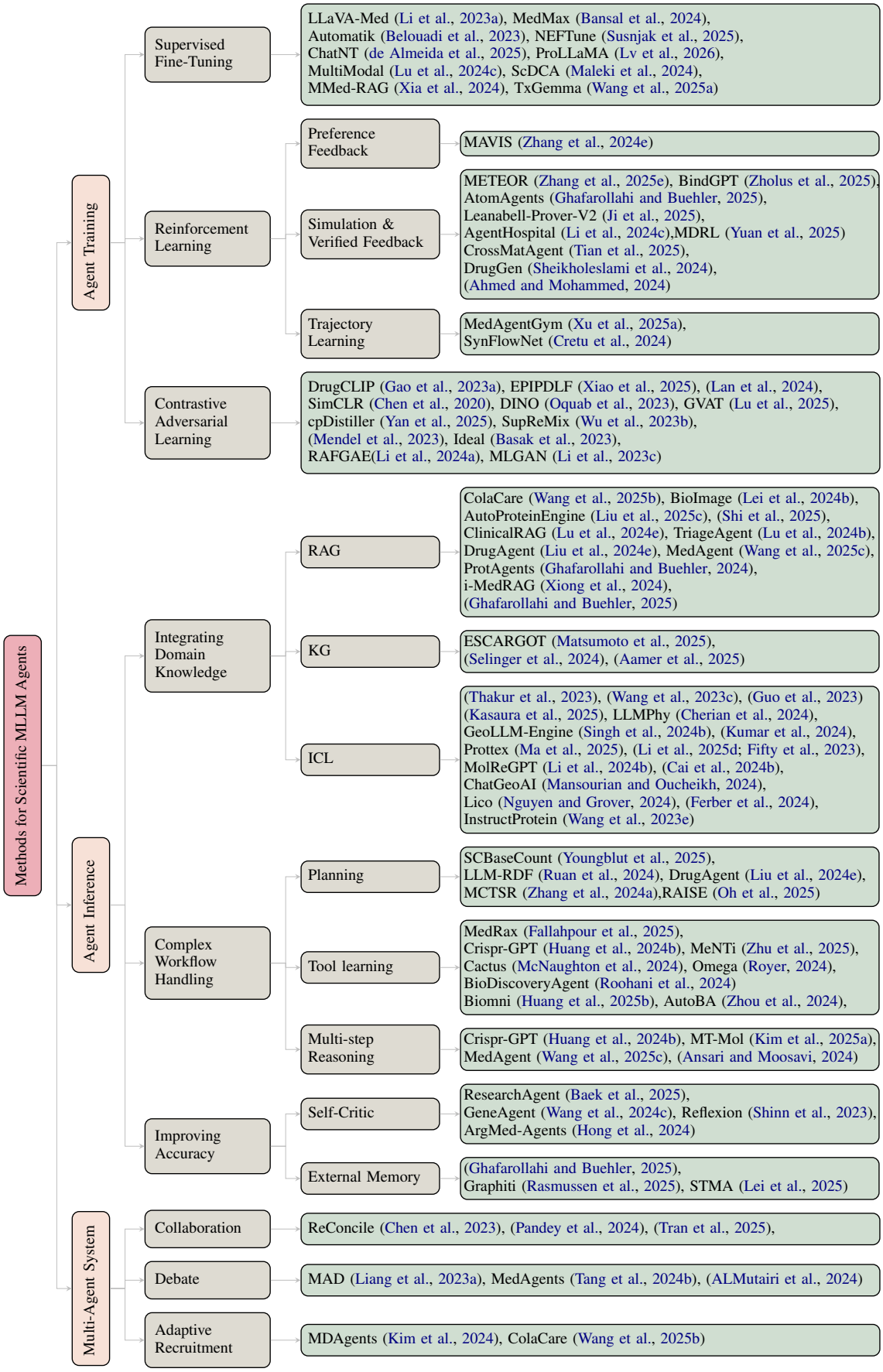


Figure 5: Taxonomy of Methods for Scientific MLLM Agents.

Abbr.	Definition	Representative Metrics
OA.	Outcome Accuracy	Acc, F1, AUC, Recall, Precision, Pass@1, Matched Score
PS.	Procedural Success	Step Correctness, Procedural Process
Suc.	Success Rate	Success Rate, Completeness, Task Completion
Sim.	Similarity	ROUGE, BLEU, BERTScore, Similarity
Err.	Error	RMSE, MAE, MAD, Error, AUC-PRC, Spread-Skill Ratio
Cal.	Calibration	Confidence, Calibration Error, Uncertainty, Entropy, Abstention Rate
Man.	Human/Expert Evaluation	Human rating, LLM-Evaluator
SV.	Scientific Validity	Validity, Novelty, Hallucination Rate
Cost	Cost	Token counts, Time

Table 6: Abbreviations used in the Metrics (MC) column of benchmark tables.

E Supplementary Benchmarks and Metrics

E.1 Comprehensive Domain Benchmarks

To provide a comprehensive evaluation landscape for scientific MLLMs, we collected a representative suite of benchmarks spanning 10 major scientific disciplines, including Biology (Table 7), Chemistry (Table 8), Physics (Table 9), Mathematics (Table 10), Environmental Science (Table 11), Geography (Table 12), Multidisciplinary Knowledge (Table 13), and General Science (Table 14). These tables detail the specific tasks the benchmark assesses, the data modalities it incorporates, its task sources, the scale (# of tasks or categories), and the primary evaluation metrics (MC) that the benchmark employs.

Task Sources indicate the origin of benchmark tasks. The abbreviations used in the tables are defined as follows:

- **Lit.**: Literature (e.g., PubMed, arXiv)
- **PubLib.**: Public datasets (e.g., Protein Data Bank, established benchmarks)
- **PriLib.**: Private datasets (e.g., in-house hospital data)
- **Syn.**: Synthetic data (e.g., software-simulated or augmented data)
- **Man.**: Human-annotated data
- **GitHub**: Online code
- **Textbk.**: Textbooks
- **Exam**: Examination problem set

Benchmark	Task	Modality					SC	#	MC
		Sq.	Sy.	V.	G.	O.			
MoleculeQA (Lu et al., 2024d)	QA	✓					PubL.	62k	OA., Err.
PGxQA (Keat et al., 2024)	QA	✓					PubL., Man.	110k	OA., HL., Man., Sim., Suc.
PubMedQA (Jin et al., 2019)	QA	✓				✓T.	Lit.	1k	OA.
MicroVQA (Burgess et al., 2025)	QA	✓		✓			Man.	1k	OA.
MedCalc-Bench (Khandekar et al., 2024)	Cal.	✓					PubL.	55	OA.
Bio-Benchmark (Jiang et al., 2025)	BI.	✓				✓T.	PubL.	1k	HL., Sim.
BioLLMBench (Sarwal et al., 2023)	BI.	✓	✓				Lit.	36	OA., Sim., Suc., Cost, Simp.
ProteinLMBench (Shen et al., 2024)	Prot.	✓					Lit., PubL.	944	OA.
LAB-Bench (Laurent et al., 2024)	Lab	✓	✓	✓		✓T.	Lit., Man., Syn.	2.4k	OA.
BioCoder (Tang et al., 2024a)	Code	✓	✓				Git., Web.	70k	OA.
BioProBench (Liu et al., 2025d)	Reas.	✓					PubL., Syn.	556k	OA., Err., Sim.
scEval (Liu et al., 2024f)	ScA.	✓					Lit.	22	OA.
SimBench (Cao et al., 2021)	ScA.					✓T.	Lit.	8 cat.	OA., Sim.
MultiMedBench (Tu et al., 2024)	BI.	✓		✓		✓T.	PubL.	14 cat.	OA., PA., Man.
DUD-E (Mysinger et al., 2012)	MD.	✓			✓	✓3D	PubL., Syn.	102	OA.
BaisBench (Luo et al., 2025a)	Omics	✓				✓T.	Lit., Syn.	229	OA.

Table 7: **Benchmarks on scientific MLLM agents (Biology).** *Note.* SC=task sources; #=test scale (cat.=categories); MC=metrics. Sq.=sequence; Sy.=symbolic; V.=vision; G.=graph; O.=other; T.=tabular. For DUD-E, Sq.=SMILES; G.=atom–bond connectivity; 3D=ligand/receptor structures. QA=question answering; BI.=biological knowledge; Reas.=protocol reasoning; ScA.=single-cell analysis; MD.=molecular docking. PubL.=public datasets; Lit.=literature; Syn.=synthetic data; Man. in SC=human-annotated data. OA.=outcome accuracy; Err./Error=prediction error; Sim.=similarity; HL.=hallucination; Suc.=success rate; Cost=token/time cost; Simp.=simplicity; Man. in MC=human/expert evaluation. Cal.=medical calculation; Prot.=protein understanding; Lab=biology lab research; Git.=GitHub; Web.=web sources; PA.=procedural accuracy. In the domain tables, Validity=scientific plausibility/testability; Novelty=departure from prior knowledge.

Benchmark	Task	Modality					Task Sources	Test Scale	Metrics
		Seq.	Symb.	Vision	Graph	Other			
PRESTO (Cao et al., 2024)	Property Prediction	✓			✓		PubLib.	878k	OA., Error, Sim., Validity
MoleculeNet (Wu et al., 2018)	Property Prediction	✓			✓	✓3D coordinates	PubLib., Syn.	700k	OA., Error
ChemBench (Mirza et al., 2024)	Property Prediction	✓					Syn., PubLib., Textbk., Exam, Lecture, Human.	2.8k	OA.
LLM4Mat-Bench (Rubungo et al., 2025)	Property Prediction	✓	✓	✓			PubLib.	2.7M	OA., Error
TOMATO-Chem (Yang et al., 2024e)	Chemical Reasoning and Design	✓	✓		✓		Lit., Human.	51	OA., Novelty, Man.
ChemLLMBench (Guo et al., 2023)	Chemical Reasoning and Design	✓		✓			PubLib.	800	OA., Man., Sim.
MaScQA (Zaki et al., 2023)	Materials Science Question Answering	✓	✓				Lit., Man.	650	OA.

Table 8: Benchmarks on scientific MLLM agents (Chemistry & Materials Science). *Note.* OA.=outcome accuracy; Error=prediction error; Sim.=similarity; Validity=scientific plausibility/testability; Novelty=departure from prior knowledge; Man.=human/expert evaluation.

Benchmark	Task	Modality					Task Sources	Test Scale	Metrics
		Seq.	Symb.	Vision	Graph	Other			
MM-PhyQA (Anand et al., 2024)	Multimodal Physics QA	✓	✓	✓			Man., PubLib., Syn.	4.5k	OA.
UGPhysics (Xu et al., 2025b)	Undergraduate Physics QA	✓	✓				Textbk., Man.	5.5k	OA.
PhysicsArena (Dai et al., 2025)	Multimodal Physics Problem Solving	✓	✓	✓			Exam, Syn., Man.	5.1k	OA., PAcc, Sim.

Table 9: Benchmarks on scientific MLLM agents (Physics)

Benchmark	Task	Modality					Task Sources	Test Scale	Metrics
		Seq.	Symb.	Vision	Graph	Other			
LeanDojo (Yang et al., 2023)	Theorem Proving in Lean	✓	✓	✓		✓Tool	PubLib., Syn.	98.7k	PSuc., Suc.
MathVerse (Zhang et al., 2024d)	Multimodal Math Problem Solving	✓	✓	✓			PubLib.	15.6k	OA., PSuc.
MathChat (Wu et al., 2023c)	Conversational Math Problem Solving	✓	✓				PubLib., Syn.	1.1k	OA., Sim., Error
GeoEval (Zhang et al., 2024b)	Geometry Problem Solving	✓	✓	✓			PubLib.	5k	OA., PSuc.
MathVista (Lu et al., 2023)	Geometry Problem Solving	✓	✓	✓			PubLib.	6.1k	OA., PSuc.
KnowledgeMath (Zhao et al., 2023)	Finance Math Knowledge Reasoning	✓	✓		✓		Exam	1.3k	OA., Man.
HARDMath (Fan et al., 2024)	Challenging Applied Math Problems	✓	✓				Exam	1.5k	OA., Error, Man.

Table 10: Benchmarks on scientific MLLM agents (Mathematics)

Benchmark	Task	Modality					Task Sources	Test Scale	Metrics
		Seq.	Symb.	Vision	Graph	Other			
FFD-IQA (Sun et al., 2023b)	Flood Disaster VQA	✓		✓			PubLib.	22k	Man.
DisasterQA (Rawat, 2024)	Disaster Response QA	✓					PubLib.	707	OA., Calib.
WeatherQA (Ma et al., 2024a)	Severe Weather VQA	✓		✓			PubLib.	8.5k	OA.
WeatherBench2 (Rasp et al., 2024)	Weather Forecasting	✓		✓		✓Spatial-grid	PubLib., Syn.	real-time updated	OA., Error, Calib.
STBench (Li et al., 2025b)	Spatio-Temporal Reasoning	✓		✓			PubLib.	70k	OA., Error
VayuBuddy (Patel et al., 2024)	Air Quality Sensor Analysis	✓	✓	✓		✓Spatial-grid	PriLib., Man.	45	OA., Suc.

Table 11: Benchmarks on scientific MLLM agents (Environment)

Benchmark	Task	Modality					Task Sources	Test Scale	Metrics
		Seq.	Symb.	Vision	Graph	Other			
EarthNets (Xiong et al., 2025b)	Remote Sensing Earth Observation			✓	✓		PubLib., GitHub, Lit., Man.	401	OA.
VLEO-Bench (Zhang and Wang, 2024)	Remote Sensing Earth Observation	✓		✓			PubLib., Lit.	2.2k	OA., Error
GEOBench (Danish et al., 2024)	Remote Sensing Earth Observation	✓		✓	✓	✓Spatial-grid	PubLib., Man., Syn.	31	OA.
Roberts et al. (Roberts et al., 2024)	Geospatial Visual Reasoning			✓		✓Temporal	Lit., Syn., Man., PubLib.	190.7k	OA., Sim.
RSUniVLM (Liu and Lian, 2024)	Remote Sensing VLM			✓		✓Temporal	PubLib., PriLib., Syn., Man., GitHub	6	OA., Sim., Error

Table 12: Benchmarks on scientific MLLM agents (Geography)

Benchmark	Task	Modality					Task Sources	Test Scale	Metrics
		Seq.	Symb.	Vision	Graph	Other			
ScienceQA (Lu et al., 2022)	Multimodal Science QA	✓		✓			PubLib., Man.	21.2k	OA., Sim., Suc.
MMSciCap+ MMSciQA (Li et al., 2024h)	Multimodal Science QA	✓		✓		✓Tabular	Lit., Syn., Man.	4k	OA., Sim.
HLE (Phan et al., 2025)	Human-Level Academic QA	✓					PriLib., Lit., Man.	2.5k	OA., Calib., Cost
MMLU (Hendrycks et al., 2020)	Multidisciplinary Knowledge QA	✓					Man., PubLib.	57	OA., Calib.
SciBench (Wang et al., 2023d)	College Scientific Reasoning	✓	✓				Lit., Man.	972	OA., Effective
GPQA (Rein et al., 2024)	Graduate-Level Expert QA	✓					Lit., Man.	448	OA., Objectivity
SuperGPQA (P Team et al., 2025)	Graduate-Level Expert QA	✓					Lit., Man.	26.5k	OA., Discrimination

Table 13: Benchmarks on scientific MLLM agents (Multidisciplinary)

Benchmark	Task	Modality					Task Sources	Test Scale	Metrics
		Seq.	Symb.	Vision	Graph	Other			
SciMON (Wang et al., 2024a)	Scientific Novelty Optimization	✓					Man., Lit.	N/A	Novelty, Error
IsoBench (Fu et al., 2024)	Multimodal Isomorphic Reasoning	✓		✓			Syn., PubLib., Lit., Man.	10	OA.
BLADE (Gu et al., 2024)	Data-Driven Scientific Analysis	✓	✓			✓Tabular	PubLib., Man.	724	OA., Calib.
FigureQA (Kahou et al., 2017)	Scientific Figure Visual QA	✓		✓		✓Tabular	Lit., Man.	15	OA., Smoothness
ArXivQA (Li et al., 2024d)	Multimodal ArXiv Paper QA	✓		✓			Man.	1k	OA., Sim.
SciCode (Tian et al., 2024)	Scientific Research Coding	✓	✓	✓		✓Tabular	Man.	338	OA., Validity
DiscoveryBench (Majumder et al., 2024)	Data-Driven Hypothesis Discovery	✓	✓				Syn., PubLib., Lit., Man., GitHub	1.2k	OA.
SciEval (Sun et al., 2023a)	Scientific Ability Evaluation	✓					PubLib., Lit., Man., Syn.	18k	OA.
MMEVALPRO (Huang et al., 2024a)	Multimodal Benchmark Calibration	✓	✓	✓			Lit.	6.4k	OA., Consistency
ScienceAgentBench (Chen et al., 2024e)	Scientific Discovery Agent Eval	✓	✓			✓Tabular	GitHub, Syn., Man., Lit.	102	Suc., Sim., Cost
Discoveryworld (Jansen et al., 2024)	Scientific Discovery Virtual Env	✓	✓			✓Tabular	Syn., Man.	120	Suc., PSuc.

Table 14: Benchmarks on scientific MLLM agents (General Science)