# Exploring Agentic Multimodal Large Language Models: A Survey for AIScientists

**Anonymous ACL submission**


## Abstract

Agentic Multimodal Large Language Models (MLLMs) are reshaping scientific discovery. These systems can perceive, reason, and act across text, images, molecules, and other data types, potentially assisting or even conducting research autonomously. This survey provides a systematic framework for building scientific MLLM agents, covering four key aspects: multimodal perception, training and inference methods, evaluation benchmarks, and human–AI collaboration. We also discuss ongoing challenges, including data scarcity, reliability, and interpretability. Rather than replacing human scientists, we view agentic MLLMs as collaborative partners that can accelerate discovery while promoting democracy and inclusivity in scientific progress. Our curated resources are available at https://anonymous.4open.science/r/awesome-agentic-AIScientists-E70B/.


## 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. Amidst this exciting revolution, scientists increasingly use AI to help tackle important research problems, from designing new drugs (Szymanski et al., 2023; M. Bran et al., 2024) and materials, to carrying out coding and model training. As a further underscore to the exciting impact of AI on scientific discovery, recent years' Nobel Prize in Chemistry[1] has also been awarded to researchers who utilized AI in pioneering new pathways for drug discovery. This recognition signals a paradigm shift, catalyzing the rapid emergence of AIScientist systems (Figure. 1).

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

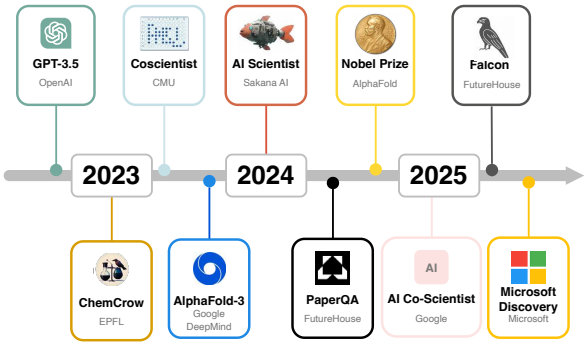


Figure 1: Timeline of Major AIScientist Systems from Leading Research Institutions.

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., 2024f), generating new hypotheses (Wang et al., 2024a; Li et al., 2024g; Baek et al., 2024a), 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., 2024e), and reviewing papers (Du et al., 2024; Weng et al., 2024).

To fully realize the potential of agentic MLLMs, it is essential to explore innovative methodologies tailored to theunique challenges within scientific domains. These challenges include cross-disciplinary knowledge integration, limited labeled data, and long-tail distributions (Wang et al., 2024c). This requires the development of advanced modeling techniques that equip models with domain knowledge, such as contrastive learning, reinforcement learning, and domain-specific tuning. Also, the deployment of these models demands robust inference strategies that empower the MLLMs to act as autonomous agents, capable of planning, self-correcting, and effectively collaborating with other agents or tools.

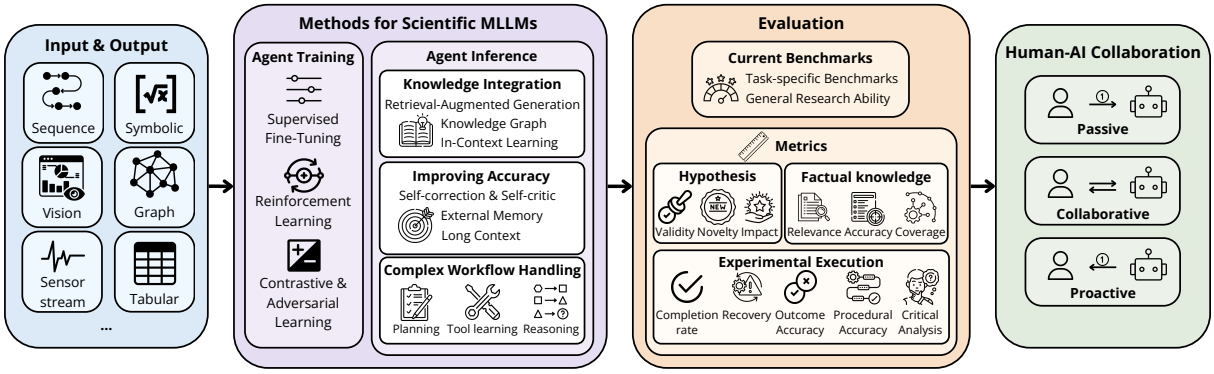


Figure 2: **Overview of the agentic MLLM framework for scientific discovery.** The pipeline includes **Input & Output** modalities (Sec 3), **Methods** covering agent training and inference (Sec 4), **Evaluation** with benchmarks and metrics (Sec 5), and **Human–AI Collaboration** (Sec 6).

As agentic MLLMs advance, new benchmarks are needed to evaluate their scientific capabilities beyond traditional metrics. Human–AI collaboration also plays a central role, ensuring that scientific agents strengthen rather than replace human expertise. This survey provides a comprehensive pipeline framework for developing agentic MLLMs in scientific contexts (Figure 2).

## 2 Scientific Workflow and the Need for Multimodal Agentic AI

Scientific discovery remains a fundamentally human pursuit based on iterative inquiry, but modern research has become increasingly complex. As shown in Fig. 3, today's scientific workflows require integrating many forms of information, from understanding current knowledge to analyzing experimental data. This need for multimodality understanding and reasoning motivates the development of the **Agentic AIScientist**. Here, we illustrate an agent built on a modular architecture that integrates Learning, Memory, Action, Cognition, and interaction with the environment (Durante et al., 2024), as shown in Fig. 3. The development of such 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 not repeated here and are provided in Appendix B.

## 3 Multimodal Perception for Scientific Agents

A core capability of agentic MLLMs is to perceive and interpret heterogeneous scientific inputs (Wang et al., 2023b; Acosta et al., 2022; Li et al., 2025e). We introduce the key data modalities and fusion mechanisms that enable cross-modal understanding in scientific agents.

### 3.1 Data Modalities

**Sequence** Sequence modeling is central to scientific discovery, spanning text (de Haan et al., 2025; Wang et al., 2023a; Bolton et al., 2024), molecules (Luo et al., 2023b; Balaji et al., 2023), and proteins (Jumper et al., 2021; Elnaggar et al., 2022; Lv et al., 2025). Foundation models leverage token representations to capture syntactic, semantic, and structural patterns (Ahmad et al., 2022; Fishman et al., 2025; Wu et al., 2024; Kim et al., 2025b). Transformer-based models such as BioMedLM (Bolton et al., 2024), GeoGalactica (Lin et al., 2023), and Meerkat (Kim et al., 2025b) capture long-range dependencies, though they often lack structural constraints addressable via graphs or 3D cues (Luo et al., 2023b; Madani et al., 2020).

**Symbolic** Symbolic equations are used to express mathematical proofs (Yang et al., 2024a), physical laws (Cherian et al., 2024; Liu et al., 2024c), and algorithms (Majumder et al., 2024; Tang et al., 2024a). Models like GPT-f and LeanCopilot (Polu and Sutskever, 2020; Song et al., 2025a) generate and verify proofs within theorem provers such as Coq and Lean (Thakur et al., 2023), while symbolic agents support code generation (Schmidgall et al., 2025) and simulation control (Li et al., 2025d; Song et al., 2023).

**Vision** Visual data such as images, videos, and spatial patterns are essential in medicine, microscopy, and geoscience (Fallahpour et al., 2025;

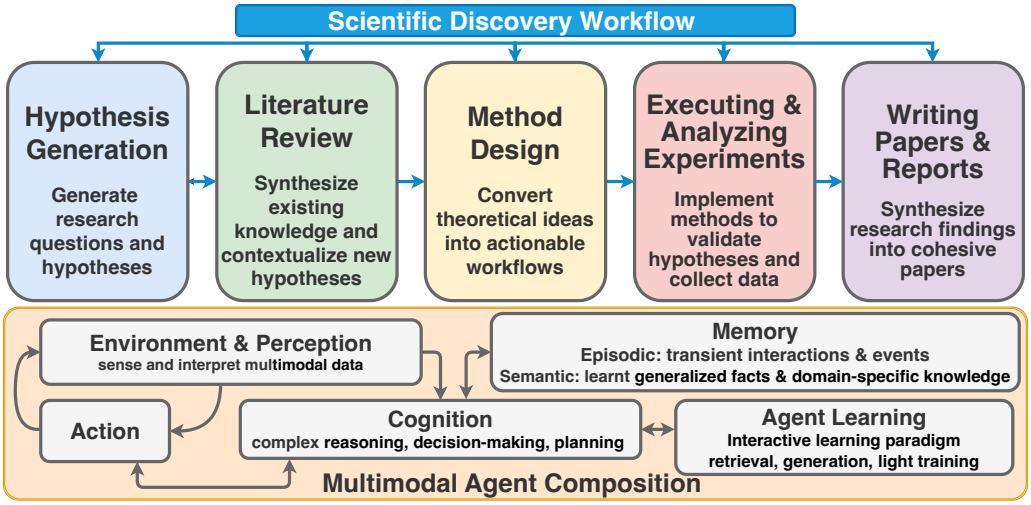


Figure 3: The Modern Scientifc Discovery Workflow and Modular Architecture of the Agentic AI System.

Zhang et al., 2025e; Yan et al., 2024). Models like BiomedCLIP and UrbanCLIP (Zhang et al., 2025e; Yan et al., 2024) extend CLIP (Radford et al., 2021) for cross-modal alignment with text and graphs (Sun et al., 2024b; Lei et al., 2024b). Diffusion models support molecule design (Zhang et al., 2025a), protein-ligand modeling (Guan et al., 2024), and medical segmentation (Qin et al., 2025). However, scientific vision remains challenging due to fine-grained reasoning demands on heterogeneous data (Chen et al., 2024a; Lei et al., 2024b).

**Graph** Graphs serve as a universal structure for representing relationships across biology (Jumper et al., 2021; Yi et al., 2022), chemistry (Gilmer et al., 2017; Reiser et al., 2022), and physics (Sanchez-Gonzalez et al., 2020), encoding structure and semantics in forms such as molecular (Liu et al., 2024d; Luo et al., 2023b) or knowledge graphs (Wang et al., 2023e; Xu et al., 2023). They also serve as structured priors for interoperable reasoning and molecule or protein function prediction (Liang et al., 2023; Abdine et al., 2024). However, integrating graphs into agentic MLLMs remains challenging due to scalability limits and the need to process evolving large-scale graphs efficiently (Feng et al., 2024; Zeng et al., 2019).

**Sensor Stream** Sensor data enables real-time analysis in biomedicine (Barman et al., 2025) and climate monitoring (Nguyen et al., 2023). Time series models like NormWear (Luo et al., 2024) align physiological signals with clinical language, while spatial-temporal models like ClimaX (Nguyen et al., 2023) handle satellite and climate data.

**Other Experimental Data** Scientific research also involves structured instrument data (Wu and Xie, 2025). **Tabular data** (e.g., scRNA-seq, proteomics) is modeled by Geneformer (Theodoris et al., 2023) and scGPT (Cui et al., 2024). **Spatial omics** combines molecular measurements with 2D coordinates (Wen et al., 2023a; Schaar et al., 2024), though high dimensionality and batch effects remain challenging.

### 3.2 Multi-modal Knowledge Fusion

Agentic MLLMs integrate diverse data types via three strategies (Tu et al., 2024; Schouten et al., 2024): **Intermediate Fusion** encodes modalities separately before concatenation, often enhanced by attention (Xu et al., 2023) or latent alignment (Zhang et al., 2024d); **Late Fusion** merges outputs from independent models (Krones et al., 2025); and **Early Fusion** combines raw inputs but requires strict alignment. Cross-modal registration remains a key bottleneck.

> **Takeaway:** Agents perceive six modalities with distinct trade-offs: **Sequences** lack structural cues; **Symbolic** requires strict syntax; **Vision** demands fine-grained annotation; **Graphs** are hard to scale; **Sensors** suffer from noise; **Tabular/Omics** need batch correction. Three fusion strategies enable cross-modal reasoning, though alignment remains challenging.

## 4 Methods for Scientific MLLM Agents

Even state-of-the-art models (e.g., GPT-4o (OpenAI, 2024)) struggle with domain-specific reasoning and tool use (Phan et al., 2025; Sawamura et al., 2025). While agentic capabilities such as reasoning, acting, and interacting primarily manifest at

inference time (Plaat et al., 2025), effective training lays the foundation for these behaviors. We consolidate methodologies across both phases, covering training that equips models with scientific knowledge, tool-use proficiency, and decision-making skills, as well as inference strategies that enable autonomous scientific workflows.

## 4.1 Agent Training

Training methods equip foundation models with capabilities essential for agentic behavior, including domain understanding via supervised learning, tool-use proficiency via function-call training (Patil et al., 2024; Qin et al., 2024), and sequential decision-making via reinforcement learning.

### 4.1.1 Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) is the dominant strategy for adapting foundation MLLMs to scientific tasks, using curated input-output datasets to update model weights (Maleki et al., 2025; Wei et al., 2022a). It has been applied to molecular property prediction (de Almeida et al., 2025), protein design (Ghafarollahi and Buehler, 2024), and therapeutic reasoning (Wang et al., 2025a). Beyond domain adaptation, SFT also enables tool-use capabilities. Gorilla (Patil et al., 2024) fine-tunes LLaMA on API documentation to generate accurate function calls, while ToolLLM (Qin et al., 2024) trains models on tool-augmented trajectories for multi-step API reasoning. Parameter-efficient methods such as LoRA (Ma, 2025) reduce costs while maintaining effectiveness in biomedical VQA (Xia et al., 2025). However, SFT alone often yields limited gains in factual correctness (Belouadi et al., 2024).

### 4.1.2 Reinforcement Learning

Reinforcement Learning (RL) is central to agentic training, enabling models to learn sequential decision-making through direct interaction with environments (Ahmed and Mohammed, 2025; AlphaProof and AlphaGeometry teams, 2024).

**Human Feedback** Human feedback is widely used to align MLLMs with human preferences in open-ended tasks such as summarization, Q&A, and personalized generation (Ouyang et al., 2022; OpenAI, 2025; Lambert et al., 2023; Sun et al., 2023b; Giuffrè et al., 2024). It is especially useful when task correctness is subjective or ill-defined. For example, ResearchAgent (Baek et al., 2025) generates ideas from academic graphs with RLHF, and MAVIS (Zhang et al., 2024c) applies DPO (Rafailov et al., 2024) using human-annotated reasoning paths (Wei et al., 2022b) to improve coherence. However, RLHF depends on manual ranking, making it costly.

**Simulation & Verified Feedback** This approach directly incorporates symbolic verification or scientific simulators into training loops (Zhou et al., 2024b; Jha et al., 2024). Formal systems such as Leanabell-Prover-V2 (Ji et al., 2025) use symbolic correctness for proof generation, while AgentHospital (Li et al., 2024d) and AtomAgents (Ghafarollahi and Buehler, 2025) simulate scientific environments for decision-making. In molecular science, models such as BindGPT (Zholus et al., 2025) and DrugGen (Sheikholeslami et al., 2024) use simulation outputs (e.g., binding scores) as reward signals. Despite high alignment fidelity, these approaches often suffer from long runtimes and high computational cost (Zhang et al., 2025g).

### 4.1.3 Contrastive & Adversarial Learning

In scientific domains with limited labels, contrastive learning (CL) enables self-supervised representation learning by pulling together paired examples and pushing apart unpaired ones (Zhou et al., 2023; Gao et al., 2023a). Strategies such as hard-negative mining (Chen et al., 2024e) and counterfactual CL (Roschewitz et al., 2025) improve generalization. Adversarial learning increases robustness by training on perturbed examples (Lan et al., 2024a; Li et al., 2024a), which is useful for distinguishing similar molecules. However, both methods depend heavily on augmentation quality and may slow or destabilize training (Xiao et al., 2025a; Salimans et al., 2016).

## 4.2 Agent Inference

Expanding test-time computation (Snell et al., 2024) offers a more efficient way to enhance outputs compared to relying on the MLLM training stage, which can be realized through diverse inference strategies (Lai et al., 2025).

### 4.2.1 Integrating Domain Knowledge

**Retrieval-Augmented Generation (RAG)** RAG improves factual accuracy by retrieving task-relevant evidence. In biomedical applications, RAG over guideline documents improves clinical alignment (Wang et al., 2025c,b; Lu et al., 2024c), while AutoProteinEngine (Liu et al., 2025e) retrieves literature to guide protein engineering.

Other systems use real-time search engines for up-to-date context (Lu et al., 2024g; Ghafarollahi and Buehler, 2024; Lei et al., 2024b). Task-specific RAG variants include i-MedRAG with query refinement (Xiong et al., 2024a), LLM Instructor with structured prompts (Liu et al., 2024e), and Hypercube-based RAG (Shi et al., 2025). However, performance is limited by retrieval corpus quality and coverage.

**Knowledge Graphs (KG)** KGs impose structured constraints for reasoning. MLLMs combined with biomedical KGs have been used for research proposal generation (Selinger et al., 2024; Aamer et al., 2025) and executable graph querying (Matsumoto et al., 2025). However, maintaining large-scale, accurate KGs remains challenging.

**In-Context Learning (ICL)** ICL is a simple and powerful strategy that adapts models to new tasks at inference by providing demonstrations, without optimizing models' parameters (Brown et al., 2020; Dong et al., 2024). It has shown strong performance across domains such as symbolic proofs (Kasaura et al., 2025; Cherian et al., 2024), geospatial analysis (Singh et al., 2024b; Mansourian and Oucheikh, 2024), medical segmentation (Kumar et al., 2024), and protein/molecule tasks (Ma et al., 2025; Li et al., 2025d; Cai et al., 2024b). ICL quality is shown to heavily depend on demonstration format, order, and similarity to the query (Liu et al., 2021a; Fung et al., 2025; Liu et al., 2024b). While adding more examples often improves accuracy (Fung et al., 2025; Wang et al., 2023d), long-context cost (Hu et al., 2025a), and degradation remain a concern (Dong et al., 2024; Bertsch et al., 2025).

#### 4.2.2 Complex Workflow Handling

**Planning** Frameworks such as ReAct, Pre-Act, and AutoAct integrate reasoning and actions for structured workflows (Yao et al., 2023; Rawat et al., 2025; Qiao et al., 2024). RAISE decomposes tasks into logical queries, while LLM-RDF combines chain-of-thought with domain tools such as Bayesian optimizers and robotic arms (Oh et al., 2025; Ruan et al., 2024). Other strategies, including Monte Carlo Tree Search (MCTS) (MarkTechPost, 2024) and biomedical LLM planners (Liu et al., 2024e), enhance planning but often face high costs.

**Tool Learning** Agents increasingly rely on domain-specific tools (Fallahpour et al., 2025; McNaughton et al., 2024; Huang et al., 2024). Early systems relied on external APIs to provide real-time access to literature, databases, and visual understanding (Bran et al., 2023; Roohani et al., 2024; Royer, 2024). Recent ecosystems such as ToolUniverse (Gao et al., 2025) and Biomni (Huang et al., 2025c) provide hundreds of scientific tools accessible via standardized protocols like MCP, enabling any LLM to orchestrate complex biomedical workflows without task-specific fine-tuning. For lab automation, agents further depend on technical documentation to interface with physical instruments (Boiko et al., 2023).

**Multi-Step Reasoning** Chain-of-thought (Chen et al., 2025b; Huang et al., 2024) and role-specialized agents improve step-wise inference. For example, MT-Mol combines RDKit for chemical reasoning (Kim et al., 2025a) for more accurate molecule synthesis design. Also, MedAgent-Pro makes medical reasoning more explainable (Wang et al., 2025c). Iterative verification also supports literature-based material property prediction (Ansari and Moosavi, 2024). However, these methods remain sensitive to early errors and limited by models' hallucinations.

#### 4.2.3 Improving Accuracy

**Self-Critique** Self-critic, also known as self-correction, refers to iterative refinement where agents evaluate and revise their own outputs (Shinn et al., 2023). Methods such as ArgMed Agents improve clinical explainability through self-argumentation iterations (Hong et al., 2024), while GeneAgent reduces hallucinations by interacting with databases and performing data analysis (Wang et al., 2024f). However, self-correction methods often increase computation.

**External Memory** Modern memory systems often adopt a two-layer structure to provide out-of-domain knowledge and reduce hallucinations, with long-term memory for persistent knowledge storage and short-term memory/tool memory (Ghafarollahi and Buehler, 2025). Temporal-aware architectures such as Graphiti (Rasmussen et al., 2025) and STMA (Lei et al., 2025) further integrate time info into memory modules. However, memory retrieval can become noisy as histories grow longer.

| Benchmark | Task | Modality | | | | | SC | # | MC |
|---|---|---|---|---|---|---|---|---|---|
| | | Sq. | Sy. | V. | G. | O. | | | |
| MoleculeQA (Lu et al., 2024f) | QA | ✓ | | | | | PubL. | 62k | OA., Err. |
| PGxQA (Keat et al., 2024) | QA | ✓ | | | | | PubL., Man. | 110k | OA., HL., Sim., ... |
| 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., Cost, ... |
| ProteinLMBench (Shen et al., 2024) | Prot. | ✓ | | | | | Lit., PubL. | 944 | OA. |
| BioCoder (Tang et al., 2023a) | Code | ✓ | ✓ | | | | Git., Web. | 70k | OA. |
| BioProBench (Liu et al., 2025f) | Reas. | ✓ | ✓ | | | ✓T. | PubL., Syn. | 556k | OA., Err., Sim. |
| scEval (Liu et al., 2024f) | ScA. | ✓ | | | | | Lit. | 22 | OA. |
| SimBench (Cao et al., 2021) | ScA. | ✓ | | | | | 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. | ✓ | | | ✓ | ✓T. | PubL., Syn. | 102 | OA. |
| BaisBench (Luo et al., 2025a) | Omics | ✓ | | | | | Lit., Syn. | 229 | OA. |

Table 1: **Benchmarks on scientific MLLM agents (Biology).** SC=Source, #=Number of tasks, MC=Metrics. Modality: Sq.=Sequence, Sy.=Symbolic, V.=Vision, G.=Graph, O.=Other. Full MC abbreviations in Appendix Tab. 6.

### 4.3 Multi-Agent System

Multi-agent systems enable complex scientific tasks by decomposing them into subproblems and coordinating solutions through strategies such as checklists and structured collaboration (Pandey et al., 2024; Tran et al., 2025). Beyond parallelization, agents can engage in multi-round discussions or debate (Liang et al., 2024; Chen et al., 2024c) to improve accuracy, using mechanisms such as confidence-weighted voting (Tang et al., 2024b; ALMutairi et al., 2024). Recent frameworks support adaptive collaboration via dynamic agent recruitment and role assignment (Kim et al., 2024; Wang et al., 2025b). However, multi-agent setups introduce new coordination challenges, including group bias and increased computational cost.

**Takeaway:** Training and inference jointly enable agentic capabilities. **SFT** provides domain knowledge but is data-hungry; **RL** enables decision-making via environment feedback but is costly; **CL** leverages unlabeled data but risks unstable training. At inference, **RAG/KG/ICL** improves reliability but depends on retrieval quality; **Planning and tool use** enables complex workflows but suffers from error accumulation; **Self-critique** enhances reasoning but increases computation. Multi-agent systems face coordination challenges.

## 5 Benchmarks and Evaluation

### 5.1 Current Benchmarks for Assessing MLLMs in Scientific Contexts

Standardized benchmarks are central to evaluating scientific MLLMs. Table 1 presents biological science benchmarks; cross-domain tables are in Appendix Tab. 7-14. Most benchmarks focus on discrete tasks like reasoning or code generation (Ronchini et al., 2022), rather than end-to-end workflows. Recent efforts such as DiscoveryWorld (Jansen et al., 2024) begin evaluating multi-step pipelines on autonomous discovery.

### 5.2 Novel Metrics Across Stages

The complexity of scientific discovery tasks calls for evaluation metrics beyond standard accuracy and efficiency (Lee et al., 2024), with criteria varying by research stage.

**Hypothesis Generation** *Validity* refers to whether a hypothesis is plausible and testable (Liu et al., 2025b; Si et al., 2024), evaluated via factual alignment, hallucination detection (Qi et al., 2024a), and feasibility (Huang et al., 2025b). *Novelty* captures departure from known knowledge. Traditional metrics (BLEU, ROUGE, BERTScore) rely on text overlap and often fail to capture scientific innovation (Wang et al., 2024b); semantic distance using scientific embeddings offers a better measure (Alkan et al., 2025). *Impact* assesses potential to open new directions (Zhang et al., 2025b), evaluated via expert assessment, citation analysis (Alkan et al., 2025), and cross-discipline comparisons (Liu et al., 2025d).

**Knowledge Retrieval** *Relevance* measures the agent's ability to surface relevant documents or visual evidence (Cai et al., 2024a; Qi et al., 2024b), using precision, recall, F1, ROUGE scores. *Coverage* measures topic breadth via citation networks or bibliographic coupling (Balasubramanian, 2025).

**Experimental Execution** *Completion/Recovery* tracks task completion rate and failure recovery (Bran et al., 2023; Cornelio and Diab, 2024). *Outcome Accuracy* uses ROC-AUC, sensitivity/specificity, and uncertainty calibration (Monaghan et al., 2021; Yang et al., 2024b). *Procedural Accuracy* evaluates correct reasoning via shortcut resistance and hallucination-aware metrics (Li et al., 2025a; Chen et al., 2024b). *Critical Analysis* assesses argument quality and causal traceability (Kargupta et al., 2025; Zhang et al., 2025c).

> **Takeaway:** Benchmarks are shifting from narrow tasks toward end-to-end discovery. **Hypothesis**: validity, novelty, and impact are key; BLEU/ROUGE often fail, prompting semantic measures; **Retrieval**: relevance and coverage are common, though broad synthesis is underexplored; **Execution**: metrics cover completion, recovery, accuracy, and procedural correctness.

## 6 The AIScientist Workflow

### 6.1 Agentic Modes of Operation

Building on HCI perspectives (Zhang et al., 2025f), we categorize agentic AI systems into three modes based on autonomy. ***Passive*** systems complete tasks autonomously after a single instruction, covering processes such as prediction and paper writing (Tang et al., 2025; Li et al., 2024h; Baek et al., 2025). ***Collaborative*** systems integrate humans into workflows, for example distributing writing tasks (Shao et al., 2025), automating protein data preprocessing (Liu et al., 2025e), or embedding checkpoints for lab approvals (Swanson et al., 2025). ***Proactive*** systems independently initiate hypotheses and execute experiments, including robotic orchestration (Tang et al., 2025; Yamada et al., 2025). All modes benefit from strategies such as multi-agent debate (Gottweis et al., 2025) and infrastructures such as LangGraph and MCP (Hou et al., 2025) to improve scalability and robustness.

### 6.2 HCI Across Research Stages

**Hypothesis & Goal Generation** This stage, often handled by agentic systems in passive or proactive modes, is constrained by limited novelty (Yamada et al., 2025) and over-reliance on existing literature (Luo et al., 2025c; Liu et al., 2025a). Tools like Owl (Skarlinski et al., 2025) help check idea novelty, while mutation (Jansen et al., 2025), crossover (Hemberg et al., 2024), and divergent strategies (Si et al., 2024) promote originality. Proactive systems further refine ideas via empirical feedback (Liu et al., 2024e) or agent-based tournaments (Gottweis et al., 2025).

**Planning** This stage transforms research goals into actionable protocols. In *passive* modes, agents auto-generate plans via existing knowledge or even literature queries (Zhou et al., 2024a; Schmidgall et al., 2025); *collaborative* systems let agents firstly propose while humans make decisions (Autoscience Institute, 2025). *Proactive* agents initiate planning, refine scope, or coordinate themselves, or even do structured multiple agent meetings (Gottweis and Natarajan, 2025; Schmidgall et al., 2025). Automated protocol generation has progressed from abstract outlines (Ghareeb et al., 2025) to detailed, executable pipelines (Zhou et al., 2024a; O'Donoghue et al., 2023), often integrating dataset specifications and model selection (Liu et al., 2025e).

**Execution** This stage covers data preparation, code generation, and self-correction. Passive agents handle retrieval, fusion, and repair (Liu et al., 2024e; Zhou et al., 2024a); collaborative ones embed domain experts or accept human guidance (Liu et al., 2024a; Jansen et al., 2025). Proactive systems retry failed steps (Royer, 2024), reflect, and invoke reviewers (Liu et al., 2024a). Evaluation uses LLM-as-a-judge (Baek et al., 2025), but it is limited by the number of tests (Lu et al., 2024a).

**Reporting** consolidates hypotheses, results, and meta-analyses into scientific documents. Passive systems draft reports based on internal validation (Jansen et al., 2025); collaborative ones include human verification (Jansen et al., 2025). Proactive agents conduct cross-experiment checks, invoke LLM-as-a-judge for peer-review simulation (Baek et al., 2025; Schmidgall et al., 2025), and apply tool-based verification as in Google Co-Scientist (Gottweis et al., 2025). Many systems adopt iterative refinement with quality checkpoints (Schmidgall et al., 2025), though limitations remain in evaluating broader scientific significance (Jansen et al., 2025).

**Resilience & Safety** are critical for AIScientists operating under uncertainty. Resilience is enabled by active learning that adjust plans based on prior results (Szymanski et al., 2023). Safety risks, such as importing unsafe code or triggering system-level failures (Lu et al., 2024a; Hamill, 2024), can be mitigated via Docker sandboxing (Zhou et al., 2024a), chemical safety checks (M. Bran et al., 2024), and standardized execution environments (Ghareeb

| Paper | Taxonomy | Ag. | DM. | Method | HCI. | Ben. | #Dom. |
|---|---|---|---|---|---|---|---|
| Zhang et al. (2024e) | Domain | ✗ | Seq.+ | Train. only | ✗ | ✓ | 6 |
| Gridach et al. (2025) | Sci. Workflow | ✓ | ✗ | Infer. only | ✓ | ✗ | 4 |
| Luo et al. (2025e) | Sci. Workflow | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Zhang et al. (2025h) | Sci. 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+ |
| **Ours** | **ML 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; Sci. Workflow=Scientific Discovery Workflow; Auto.=Autonomy Level.

et al., 2025). These safeguards allow systems to recover from failure while ensuring reproducibility and preventing harm (Tang et al., 2025).

**Takeaway:** AIScientists operate in three modes. **Passive**: agents complete tasks after a single human instruction; **Collaborative**: humans are embedded via checkpoints and approvals; **Proactive**: agents independently initiate hypotheses and experiments. Each mode presents distinct tradeoffs in human-AI collaboration, with shared practices emerging across research stages. Currently, collaborative modes remain mainstream, as they balance AI capabilities with human expertise. Safety mechanisms (sandboxing, chemical checks) help ensure reproducibility and prevent harm.

## 7 Related Survey

Several surveys have examined scientific AI agents from different angles: workflow stages (Luo et al., 2025e; Zhang et al., 2025h), autonomy levels (Gridach et al., 2025; Ren et al., 2025), and domain resources (Wei et al., 2025; Hu et al., 2025b). Our survey complements these with a **pipeline-oriented** perspective, analyzing data modalities (Sec. 3), training and inference methods (Sec. 4), benchmarks (Sec. 5), and human-AI collaboration (Sec. 6), aiming to clarify not only *what* agents exist, but *how* they are built.

## 8 Future Work

**Multimodal Perception** Agentic MLLMs demonstrate remarkable flexibility in processing diverse scientific data types, from sequences and graphs to sensor streams and omics. This versatility positions them as powerful tools for cross-disciplinary research. However, challenges remain: sequences lack structural cues, graphs struggle with scalability, sensors suffer from noise, and tabular data require batch correction. Advancing cross-modal alignment will be essential for seamless integration of heterogeneous scientific inputs.

**Training and Inference** We have entered an era where post-training and inference-time strategies unlock powerful intelligence for various tasks in agentic Scientists. These systems can now reason, plan, and self-correct within complex scientific workflows. Yet challenges still persist. Some tasks lack sufficient data for effective SFT. RL remains computationally expensive, and contrastive learning risks instability. Exploring how agents can improve performance under few-shot settings or guided by scientific priors represents a promising direction.

**Evaluation** Benchmarks are increasingly moving beyond isolated task metrics toward end-to-end workflows, enabling agents to function as genuine research assistants. In this context, evaluation extends beyond outcome accuracy to encompass scientific considerations such as novelty, hypothesis impact, and reproducibility. Designing comprehensive and effective evaluation protocols that capture these dimensions remains an important open challenge.

**Human-AI Collaboration** While collaborative modes currently dominate, balancing agent autonomy with human oversight remains challenging. Future systems should integrate adaptive checkpoints, trust calibration, and safety mechanisms (e.g., sandboxing, chemical checks) to ensure reproducibility and prevent harm.

**Toward More Adaptive Agents** Future agents may need to adapt their pipelines to new domains with minimal human intervention. While current systems require manual configuration, future agents may learn to compose components dynamically based on task specifications.

## Limitations

**Data Modality** Aligning heterogeneous scientific data remains a core challenge due to the scarcity of high-quality, cross-modal datasets (Zhang et al., 2025d; Reddy and Shojaee, 2024; Acosta et al., 2022). Scientific sensing data further complicate modeling with issues of sparse sampling (Schouten et al., 2024). In addition, concerns over privacy, licensing, and reproducibility hinder dataset sharing, limiting progress on open benchmarks and unified frameworks (Liu et al., 2025c).

**Agent Methodology** Scientific MLLM Agents face challenges beyond passive, general-purpose MLLMs. The first is verifiability: agents often generate chemically invalid molecules or non-causal claims without proper checks (Gridach et al., 2025; Smbatyan et al., 2025). The second is lab integration: real-world deployment depends on interfacing with fragmented APIs and incomplete metadata, where errors such as unit miscalibration pose serious technical and safety risks (Hastings, 2023). These limitations call for agents that incorporate domain knowledge, enforce scientific constraints, and integrate reliably with experimental systems.

**Benchmarks** Current benchmarks for AIScientists remain limited in modality, metrics, and task realism (Chen et al., 2025c; Tang et al., 2025). Most focus narrowly on text and vision, overlooking critical modalities including sensor data, simulations, and experimental logs (Barman et al., 2025; Mo and Liang, 2024). Standard metrics (e.g., accuracy, BLEU, ROUGE) fall short in evaluating novelty (Wang et al., 2024b), reproducibility, or long-term scientific impact (Majumder et al., 2024). High-level skills, such as hypothesis generation and experimental planning, still rely on expert judgment, as automated metrics lack robustness and consensus (Keat et al., 2024; Yang et al., 2025b). Moreover, most benchmarks test isolated tasks rather than full workflows or human-AI collaboration (Jansen et al., 2024). These gaps limit our ability to assess AIScientists' real-world capabilities.

**Human Expertise** Human-AIScientist collaboration spans passive, collaborative, and proactive modes, but all are prone to limited situational awareness (Shao et al., 2025). Agents often misinterpret human intent or overlook experimental constraints, disrupting workflows or yielding suboptimal results (Luo et al., 2025d). Also, in collaborative settings, frequent human approvals can slow progress and cause human fatigue, while too few checkpoints risk trust and safety (Swanson et al., 2025; Autoscience Institute, 2025). Meanwhile, mechanisms such as sandboxing and human overrides improve safety, but rigid controls can limit models' adaptability and exploration (Zhou et al., 2024a; Tang et al., 2025), posing a key design dilemma for resilient human-AIScientist interaction (Yadav and Shahi, 2024; M. Bran et al., 2024).

## References


Naafey Aamer, Muhammad Nabeel Asim, Shan Munir, and Andreas Dengel. 2025. Automating ai discovery for biomedicine through knowledge graphs and llm agents. *bioRxiv*, pages 2025–05.

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. 2025. 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. 2025. Improving targeted molecule generation through language model fine-tuning via reinforcement learning. *arXiv preprint arXiv:2405.06836*.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G Iyer, et al. 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 *Pacific-Asia Conference on 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.

Autoscience Institute. 2025. Meet carl: The first AI system to produce peer-reviewed research. Blog post. Accessed: 2025-03-06.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. 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. 2024a. Researchagent: Iterative research idea generation over scientific literature with large language models. *ArXiv*, abs/2404.07738.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024b. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

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.*

Hritik Bansal, Daniel Israel, Siyan Zhao, Shufan Li, Tung Nguyen, and Aditya Grover. 2025. Medmax: Mixed-modal instruction tuning for training biomedical assistants. *Computing Research Repository*, 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, et al. 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. 2023a. 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.

Hritam Basak, Soumitri Chattopadhyay, Rohit Kundu, Sayan Nag, and Rammohan Mallipeddi. 2023b. 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. 2024. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. 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, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,

Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 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, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. 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*, abs/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, Kejie Li, Nicolas Müller, Dario Righelli, Carmel Regev, Wolfgang Huber, Michael Hawrylycz, Mingyao Li, Xiaowei Zhan, Wolfgang Huber, John Marioni, and Stephanie C. Hicks. 2021. A benchmark study of simulation methods for single-cell rna sequencing data. *Nature Communications*, 12:6911.

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*, abs/2406.10185.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024c. 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, et al. 2025b. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

Z Chen, M Varma, JB Delbrouck, et al. 2024d. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *Preprint at https://arxiv.org/abs/2401.12208*.

Zhengyi Chen, Chunmin Zhang, Biyun Zhang, and Yifan He. 2024e. 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.

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. 2025c. 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. 2023. 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. 2025. 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. 2025. 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*.

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. 2024. 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, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 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, Zhenghua Chen, Min Wu, Xiaoli Li, Jay Lee, et al. 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 & 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 *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. 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*.

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, et al. 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*.

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, et al. 2023b. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Yue Gao, Kunlun Zhu, Kexin Huang, and Marinka Zitnik. 2025. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*.

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

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 and Vivek Natarajan. 2025. Accelerating scientific breakthroughs with an ai co-scientist. *Google Research Blog*.

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

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*.

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

Jasper Hamill. 2024. Ai scientist llm goes rogue: Creators warn of "significant risks" and "safety concerns". `https://www.thestack.technology/ai-scientist-llm-goes-rogue/`. Accessed: 2025-10-29.

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.

J. Hastings. 2023. Ai for scientific discovery.

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*.

Erik Hemberg, Stephen Moskal, and Una-May O'Reilly. 2024. Evolving code with a large language model. *arXiv preprint arXiv:2401.07102*.

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, Lulu Chen, Fengxiang Wang, Yuanyuan Zhang, Xiangyu Zhao, Feilong Tang, Encheng Su, Junzhi Ning, Xinyao Liu, Ye Du, Changkai Ji, Cheng Tang, Huihui Xu, Ziyang Chen, Ziyan Huang, Jiyao Liu, Pengfei Jiang, Yizhou Wang, Chen Tang, Jianyu Wu, Yuchen Ren, Siyuan Yan, Zhonghua Wang, Zhongxing Xu, Shiyan Su, Shangquan Sun, Runkai Zhao, Zhisheng Zhang, Yu Liu, Fudi Wang, Yuanfeng Ji, Yanzhou Su, Hongming Shan, Chunmei Feng, Jiahao Xu, Jiangtao Yan, Wenhao Tang, Diping Song, Lihao Liu, Yanyan Huang, Lequan Yu, Bin Fu, Shujun Wang, Xiaomeng Li, Xiaowei Hu, Yun Gu, Ben Fei, Zhongying Deng, Benyou Wang, Yuewen Cao, Minjie Shen, Haodong Duan, Jie Xu, Yirong Chen, Fang Yan, Hongxia Hao, Jielan Li, Jiajun Du, Yanbo Wang, Imran Razzak, Chi Zhang, Lijun Wu, Conghui He, Zhaohui Lu, Jinhai Huang, Yihao Liu, Fenghua Ling, Yuqiang Li, Aoran Wang, Qihao Zheng, Nanqing Dong, Tianfan Fu, Dongzhan Zhou, Yan Lu, Wenlong Zhang, Jin Ye, Jianfei Cai, Wanli Ouyang, Yu Qiao, Zongyuan Ge, Shixiang Tang, Junjun He, Chunfeng Song, Lei Bai, and Bowen Zhou. 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. 2025a. 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. 2024. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *Computing Research Repository*, arXiv:2404.18021.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. 2025b. Automated hypothesis validation with agentic sequential falsifications. *ArXiv*, abs/2502.09858.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Junze Zhang, Yin Di, Shruti Marwaha, Jennefer Carter, Xin Zhou, Matthew T Wheeler, Jon Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. 2025c. Biomni: A general-purpose biomedical ai agent. *bioRxiv*.

Qian Huang, Hongyu Ren, Peng Chen, Gregor Kržmanc, Daniel Zeng, Percy S Liang, and Jure Leskovec. 2023a. Prodigy: Enabling in-context learning over graphs. *Advances in Neural Information Processing Systems*, 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*, 37:10088–10116.

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, et al. 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, et al. 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, et al. 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. 2018. 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*, abs/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, et al. 2024. Pgxqa: A resource for evaluating llm performance for pharmacogenomic qa tasks. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, 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, et al. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

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

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, et al. 2025b. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won 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).

Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, et al. 2025. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791.

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. 2024a. Generating mutants of monotone affinity towards stronger protein complexes through adversarial learning. *Nature Machine Intelligence*, 6(3):315–325.

Tian Lan, Shuquan Su, Pengyao Ping, Gyorgy Hutvagner, Tao Liu, Yi Pan, and Jinyan Li. 2024b. 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. 2024. 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, et al. 2024a. Alzheimer's disease diagnosis from multi-modal 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, et al. 2023. 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328.

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*, 36:28541–28564.

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.

Guanghui Li, Shuwen Li, Cheng Liang, Qiu Xiao, and Jiawei Luo. 2024b. 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*, abs/2504.12314.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024c. 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, et al. 2024d. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

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

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024f. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Annual Meeting of the Association for Computational Linguistics*.

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

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

Ning Li, Liang Cheng, Lei Wang, Hui Chen, Yalu Zhang, Yunchang Yao, Manchun Li, et al. 2024i. 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.

Rui Li, Andrew Wen, Jing Gao, and Hongfang Liu. 2023d. 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. 2025e. 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. 2024. 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. 2023. Drugchat: towards enabling chatgpt-like capabilities on drug molecule graphs. *arXiv preprint arXiv:2309.03907*.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, et al. 2023. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

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

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. 2025b. Hypobench: Towards systematic and principled benchmarking for hypothesis generation. *ArXiv*, abs/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. *Machine Learning Repository*, 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*.

Xianyuan Liu, Jiayang Zhang, Shuo Zhou, Thijs L. van der Plas, Avish Vijayaraghavan, Anastasiia Grishina, Mengdie Zhuang, Daniel Schofield, Christopher Tomlinson, Yuhan Wang, Ruizhe Li, Louisa van Zeeland, Sina Tabakhi, Cyndie Demeocq, Xiang Li, Arunav Das, Orlando Timmerman, Thomas Baldwin-McDonald, Jinge Wu, Peizhen Bai, Zahraa Al Sahili, Omnia Alwazzan, Thao N. Do, Mohammod N. I. Suvon, Angeline Wang, Lucia Cipolina-Kun, Luigi A. Moretti, Lucas Farndale, Nitisha Jain, Natalia Efremova, Yan Ge, Marta Varela, Hak-Keung Lam, Oya Celiktutan, Ben R. Evans, Alejandro Coca-Castro, Honghan Wu, Zahraa S. Abdallah, Chen Chen, Valentin Danchev, Nataliya Tkachenko, Lei Lu, Tingting Zhu, Gregory G. Slabaugh, Roger K. Moore, William K. Cheung, Peter H. Charlton, and Haiping Lu. 2025c. Towards deployment-centric multimodal ai beyond vision and language. *arXiv preprint arXiv:2504.03603*.

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. 2025d. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *ArXiv*, abs/2503.21248.

Yungeng Liu, Zan Chen, Yuguang Wang, and Yiqing Shen. 2025e. 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. 2025f. 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. *Computing Research Repository*, arXiv:2408.06292.

Meng Lu, Ho Brandon, Ren Dennis, and Xuan Wang. 2024b. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *ICML 2024 AI for Science Workshop*.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024c. 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, et al. 2024d. 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. 2024e. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024f. 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. 2024g. 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.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024h. 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.

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. 2025b. Free: The foundational

semantic recognition for modeling environmental ecosystems. *arXiv preprint arXiv:2311.10255*.

Yi Luo, Linghang Shi, Yihao Li, Aobo Zhuang, Yeyun Gong, Ling Liu, and Chen Lin. 2025c. From intention to implementation: automating biomedical research via llms. *Science China Information Sciences*, 68(7):1–18.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023b. 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, Atoosa Kasirzadeh, and Nihar B. Shah. 2025d. The more you automate, the less you see: Hidden pitfalls of ai scientist systems. *arXiv preprint arXiv:2509.08713*.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025e. 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. 2025. Prollama: A protein large language model for multi-task protein language processing. *IEEE Transactions on Artificial Intelligence*.

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*, pages 1–11.

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*.

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*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 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. 2025. Efficient fine-tuning of single-cell foundation models enables zero-shot molecular perturbation prediction. *arXiv preprint arXiv:2412.13478*.

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 2024, Vol. 13, Page 348*, 13:348.

MarkTechPost. 2024. Enhancing mathematical reasoning in llms: Integrating monte carlo tree search with self-refinement. https://www.marktechpost.com/2024/06/18/enhancing-mathematical-reasoning-in-llms-integrating-monte-carlo-tree-search-with-self-refinement/. Accessed: 2025-08-25.

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. 2023a. Exploring the effects of contrastive learning on homogeneous medical image data. In *BVM Workshop*, pages 128–133. Springer.

Robert Mendel, David Rauber, and Christoph Palm. 2023b. Exploring the effects of contrastive learning on homogeneous medical image data. In *BVM Workshop*, 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, et al. 2024. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

Shentong Mo and Paul Pu Liang. 2024. Multimed: Massively multimodal and multitask medical understanding. *arXiv preprint arXiv:2408.12682*.

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.

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.

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*.

Odhran O'Donoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Essa Ghareeb, Justin Booth, and Samuel G Rodriques. 2023. Bioplanner: Automatic evaluation of llms on protocol planning in biology. *arXiv preprint arXiv:2310.10632*.

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. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2025. 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, et al. 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*.

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 *Forty-first International Conference on Machine Learning*.

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 G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2024. Gorilla: Large language model connected with massive apis. In *International Conference on Learning Representations*.

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, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco

Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peri-

styy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu

Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Aske Plaat, Walter Kosters, and Mike Preuss. 2025. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037*.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *CoRR*, abs/2009.03393.

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 *Natural Language Processing and Chinese Computing*.

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, et al. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*.

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*, abs/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. 2024. 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, et al. 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*.

Chandan K Reddy and Parshin Shojaee. 2024. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. *arXiv preprint arXiv:2412.11427*.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. 2023. 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, et al. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 554–563.

S. Ronchini, M. Branchesi, G. Oganesyan, B. Banerjee, U. Dupletsa, G. Ghirlanda, J. Harms, M. Mapelli, and F. Santoliquido. 2022. Perspectives for multi-messenger astronomy with the next generation of gravitational-wave detectors and high-energy satellites. *Astronomy & Astrophysics*, 665:A97.

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*, page 103668.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Tabula Sapienrosens Consortium, Stephen R Quake, and Jure Leskovec. 2023. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pages 2023–11.

Loïc A Royer. 2024. Omega—harnessing the power of large language models for bioimage analysis. *Nature Methods*, pages 1–3.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. 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, Peter Norvig, and Artificial Intelligence. 1995. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25(27):79–80.

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*.

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, et al. 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*, pages 2024–12.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. 2025. 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. 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. 2025a. 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, et al. 2025b. 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. corr, abs/2306.06624, 2023. doi: 10.48550. *arXiv preprint arXiv.2306.06624*.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024a. Scieval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*.

Yimin Sun, Chao Wang, and Yan Peng. 2023a. 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.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. 2024b. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. *arXiv preprint arXiv:2305.15072*.

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. 2023b. 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*, pages 1–3.

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

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

Xiangru Tang, Bill Qian, Rick Gao, Jiakang Chen, Xinyun Chen, and Mark Gerstein. 2023a. Biocoder: A benchmark for bioinformatics code generation with contextual pragmatic knowledge.

Xiangru Tang, Bill Qian, Rick Gao, Jiakang Chen, Xinyun Chen, and Mark Gerstein. 2024a. Biocoder: A benchmark for bioinformatics code generation with large language models. *arXiv preprint arXiv:2308.16458*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023b. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

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.

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, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

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, et al. 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 Dylan 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, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. 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, et al. 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, et al. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, et al. 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. *Artificial Intelligence Repository*, 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, et al. 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, Doug Downey, Heng Ji, and Tom Hope. 2024b. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 279–299. Association for Computational Linguistics.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. 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.

Qingyun Wang, Zixuan Zhang, Hongxiang Li, Xuan Liu, Jiawei Han, Huimin Zhao, and Heng Ji. 2024c. Chem-FINESE: Validating fine-grained few-shot entity extraction through text reconstruction. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1–16, St. Julian's, Malta. 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*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024d. 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. 2024e. Autosurvey: Large language models can automatically write surveys. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2023d. 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. 2024f. Geneagent: Self-verification language agent for gene set knowledge discovery using domain databases. *Artificial Intelligence Repository*, arXiv:2405.16205.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala, and Rishita Anubhai. 2023e. 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, page 2250–2261, New York, NY, USA. Association for Computing Machinery.

Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. 2025c. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *Artificial Intelligence Repository*, 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. 2022a. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 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, Siqi Sun, Lei Bai, and Bowen Zhou. 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*, 35:4913–4927.

Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. 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. 2023a. 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. 2023b. 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.

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, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

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. 2025a. 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. 2025b. 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. 2023b. 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*.

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. 2025. 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, et al. 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. 2025a. Epipdlf: a pretrained deep learning framework for predicting enhancer–promoter interactions. *Bioinformatics*, 41(5):btae716.

Zhichao Xiao, Yan Li, Yijie Ding, and Liang Yu. 2025b. 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. 2024a. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

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

Zhitong Xiong, Fahong Zhang, Yi Wang, Yilei Shi, and Xiao Xiang Zhu. 2024b. Earthnets: Empowering artificial intelligence for earth observation. *IEEE Geoscience and Remote Sensing Magazine*.

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, et al. 2025a. Medagentgym: Training llm agents for code-based medical reasoning at scale. *Computation and Language Repository*, 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. *Int. J. Trop. Med*, 19:92–98.

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, et al. 2025a. Triple-effect correction for cell painting data with contrastive and domain-adversarial learning. *Nature Communications*, 16(1):6886.

Chengwei Yan, Yu Zhang, Jiuxin Feng, Heyang Hua, Zhihan Ruan, Zhen Li, Siyu Li, Chaoyang Yan, Pingjing Li, Jian Liu, et al. 2025b. 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. 2025a. 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. 2024a. 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*, 36:21573–21612.

Ruixin Yang, Dheeraj Rajagopal, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. 2024b. Confidence calibration and rationalization for llms via multi-agent deliberation. *ArXiv*, abs/2404.09127.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, et al. 2024c. 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. 2025b. 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. 2023. 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, et al. 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, et al. 2025. scbasecount: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv*, pages 2025–02.

Botao Yu, Frazier N. Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. 2025. 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, NM Krishnan, et al. 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. 2025a. An optimized multi-task contrastive learning framework for hifu lesion detection and segmentation. *Scientific Reports*, 15(1):29666.

Matineh Zavar, Hamid Reza Ghaffari, and Hamid Tabatabaee. 2025b. 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 K. Prasanna. 2019. Graphsaint: Graph sampling based inductive learning method. *CoRR*, abs/1907.04931.

Chenhui Zhang and Sherrie Wang. 2024. Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*.

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, Marwin Segler, Maik Riechert, Li Yuan, Lei Chen, Haiguang Liu, and Tao Qin. 2025a. Unigenx: Unified generation of sequence and structure with autoregressive diffusion. *arXiv preprint arXiv:2503.06687*.

Haoxuan Zhang, Ruochi Li, Yang Zhang, Ting Xiao, Jiangping Chen, Junhua Ding, and Haihua Chen. 2025b. The evolving role of large language models in scientific innovation: Evaluator, collaborator, and scientist. *arXiv preprint arXiv:2507.11810*.

Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Xinyu Zhang, Fangzhi Xu, Qika Lin, Rui Mao, Erik Cambria, and Jun Liu. 2025c. 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. 2024a. Geoeval: Benchmark for evaluating llms and multi-modal models on geometry problem-solving. *arXiv preprint arXiv:2402.10104*.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2025d. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, 57(6):1–38.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. 2024c. 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, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2024d. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

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, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2025e. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640.

Shuning Zhang, Hui Wang, and Xin Yi. 2025f. 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. 2025g. Generation of rational drug-like molecular structures through a multiple-objective reinforcement learning framework. *Molecules*, 30(1):18. Copyright - © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2025-01-10.

Yanbo Zhang, Sumeer A Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, et al. 2025h. 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. 2024e. 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. *CoRR*.

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, et al. 2024a. 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. 2024b. 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, et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163.

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.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2024. Menti: Bridging medical calculator and llm agent with nested tool calling. *arXiv preprint arXiv:2410.13610*.

# APPENDIX

## A Released Artifacts

Github:
https://anonymous.4open.science/r/awesome-agentic-AIScientists-E70B/

## 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 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., 2025a). 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., 2025e; Ferrag et al., 2025). It also involves subtle, collaborative behaviors that are hard to formalize but central to scientific creativity (Agarwal et al., 2025; 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., 2025).

**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., 2025e). 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). Also, the integration of semantic web search and citation mapping further ensures the resulting literature reviews are both comprehensive and up-to-date, bridging gaps that have limited manually literature synthesis (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 conduct experiments, simulations, or laboratory testing of the designed methods

(Gridach et al., 2025; Eger et al., 2025). Then, they rely on the systematic gathering of 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., 2025)), public repositories (e.g., multiomics data (Nam et al., 2024), genomic sequences (Liu et al., 2024a), optical remote sensing (Li et al., 2024i)), 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., 2025e; 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 outcome a dramatic improvement in research productivity, where repetitive and time-intensive tasks become automated and researchers are allowed to focus on high-level conceptual challenges (Eger et al., 2025). Additionally, the use of such systems strengthens scientific rigor by providing consistent and reproducible outputs that minimize human error and subjectivity (Gridach et al., 2025). Forwards, the continual evolution of these systems promises to redefine the limits of scientific research cycles and fuel an era of innovation and discovery.

### 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 et al., 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 typically built on large pre-trained foundation models, typically 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 collec-

tively 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 literature, 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., 2025) 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., 2019), ChemCrow (Bran et al., 2023), AIScientist(Lu et al., 2024a), and Biomni (Huang et al., 2025c) 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., 2025b), and ORGANA (Darvish et al., 2025)). 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., 2022b), 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.

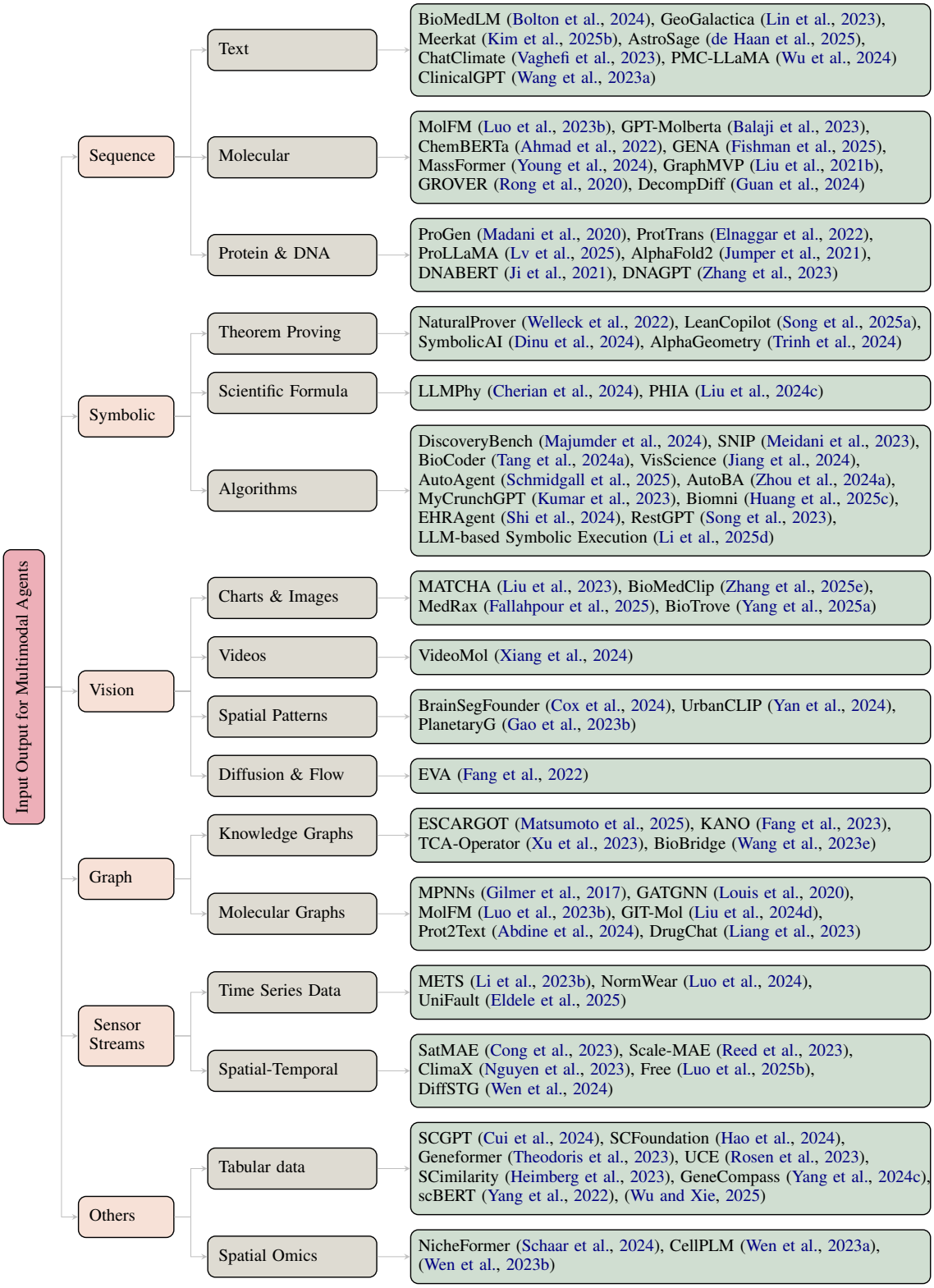


Figure 4: Taxonomy of Data Modality for Scientific MLLM Agents.

# 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., 2024d; Ma et al., 2024b; Lei et al., 2024b; Chen et al., 2024d; Royer, 2024; Lu et al., 2024d; 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., 2023), Spatial omics (Schaar et al., 2024; Wen et al., 2023a,b).

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., 2024d). 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.

| Feature | Early Fusion | Intermediate Fusion | Late Fusion |
|---|---|---|---|
| **Fusion Timing** | Before Encoding (Raw Input Layer) | Between Encoder and Prediction Layer | After Model Output |
| **Operation** | Raw-input Alignment | Simple Concatenation, Attention, Latent-space Alignment | Prediction Averaging, Traditional ML |
| **Cross-Modal Interaction** | Deepest | Medium to High | Weakest |
| **Challenge** | Strict Requirement | Designing Efficient Fusion Modules | Lacks Deep Inter-modality Interaction |
| **Computational Complexity** | Usually High (Requires training a single large encoder) | Medium | Low (Independent Encoders, Simple Prediction Integration) |
| **Examples** | DMGNN (Lei et al., 2024a), KANO (Fang et al., 2023) | TCA-Operator (Xu et al., 2023), BiomedCLIP (Zhang et al., 2024d), CAMR (Wu et al., 2023a) | (Jung et al., 2024) |

Table 3: Comparison of Multimodal Knowledge Fusion Strategies.

# 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, and limitations 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, while keeping costs manageable (several hundred dollars or less) (Maleki et al., 2025; Wei et al., 2022a). Thus, supervised fine-tuning (SFT) has become the most prevalent strategy for customizing foundation models to specialized tasks (Xu et al., 2025a). Here, we introduced 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, common material can be biomedical image-caption annotation (Li et al., 2023a; Bansal et al., 2025). 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., 2024; 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 (Ghafarollahi and Buehler, 2024), 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., 2025; Lu et al., 2024d).

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., 2025) incorporated a drug-conditional adapter into scGPT (a pretrained single-cell sequence foundation model), and MMed-RAG (Xia et al., 2025) applied LoRA adapters to align 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

| Strategy | Mechanism | Advantages | Limitations | Scientific Examples |
|---|---|---|---|---|
| **Supervised Fine-Tuning (SFT)** | Updates model weights using input-output data pairs. | Cost-effective than re-training models and dominant for task-specific adaptation; Simpler implementation. | Highly data-dependent & yields limited gains in factual correctness. | Molecular property prediction, Protein design, Biomedical VQA. |
| **Reinforcement Learning (RL)** | Optimizes sequential decision by maximizing rewards from an environment. | Provides structured feedback (human, simulation, trajectory). | High computational cost and probably long run-times; expensive on RLHF. | RLHF: Research idea generation; Simulation: Drug discovery, Formal proof generation. |
| **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. |
| **Adversarial Learning (AL)** | Trains a Generator and a Discriminator in an adversarial process. | Significantly improves robustness and generalization in noisy or ambiguous scientific data. | Can slow down training; Risks unstable training and failure to converge. | Distinguishing chemically similar molecules. |

Table 4: Comparison of Scientific MLLM Agent Training Strategies.

| Strategy | Mechanism | Advantages | Limitations | Scientific Examples |
|---|---|---|---|---|
| **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 cause long-context cost. | **RAG:** Clinical guideline adherence; **KG:** Research proposal planning; **ICL:** Few-shot learning for symbolic proofs. |
| **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 softwares or lab instruments. |
| **Improving Accuracy** | Enables agents to evaluate its own output 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. |
| **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. |

Table 5: Comparison of Scientific MLLM Agent Inference Strategies.

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., 2024; Maleki et al., 2025). 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 great potential 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, 2025; Sheikholeslami et al., 2024). RL requires different rewards, such as expert preferences, simulation, synthetic reward, and agents' evaluation.

**Human Feedback** Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) is widely used to directly align MLLMs with human preferences and human values, such as in ChatGPT (OpenAI, 2025; Lambert et al., 2023; Sun et al., 2023b; Giuffrè et al., 2024). RLHF is particularly useful for tasks where the "correct" answer is subjective or difficult to define formally. For instance, RLHF enables models to better reflect diverse human opinions in open-ended tasks such as summarization, Q&A, or generating personalized responses (Baek et al., 2025; Zhang et al., 2024c). For example, some agentic scientists (Baek et al., 2025) generated research ideas with academic graphs aligned with RLHF. Also, MAVIS (Zhang et al., 2024c) utilized DPO (Rafailov et al., 2024) on the human-annotated chain-of-thought reasoning path (Wei et al., 2022b) to refine mathematical logical coherence. However, RLHF requires human preference ranking, making it expensive and hard to scale.

**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., 2024b; 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., 2024d). 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, 2025) generate drug-like molecules, simulate binding or stability scores, and use them as reward signals for further refinement. While simulation feedback offers highly structured and reliable signals for model alignment, it faces challenges such as high computational cost and long run-times, particularly in physically intensive domains (Zhang et al., 2025g; 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., 2025). However, this method is underexplored in agentic AIScientists, and the construction cost is high.

#### D.2.3 Contrastive & Adversarial Learning

In the scientific domain, datasets with labels are expensive to obtain (Roschewitz et al., 2025; Xiao et al., 2025a; Huang et al., 2023b). Contrastive learning is a kind of self-learning strategy, which used unlabeled data to learn representations (Zhou et al., 2023; Azizi et al., 2023; Chen et al., 2020). The key points are to form a dataset that contains pair-wise positive examples, while recognizing other unpaired data as negative examples, and we can build a large-scale dataset without much cost. For example, protein-molecule pair dataset (Gao et al., 2023a), or enhancers and promoters interaction on DNAs (Xiao et al., 2025a).

Contrastive learning (CL) framework starts with creating embedding for multiple different modality and then creates a matrix that is used to calculate pair similarity (e.g., cross-entropy (Xiao et al., 2025a), nearest neighbor relationship (Yan et al., 2025b)), dot-product (Li et al., 2021)). In this way, CL teaches a model to bring positive examples closer in its representation space while pushing negative ones further apart, to learn the representation. For building positive pairs, there are many different methods: SimCLR is the simplest, effective, and widespread method in image representation learning, which uses photometric and geometric transformations to the original image (Zavar et al., 2025b; Chen et al., 2020). Also, some useful strategy merges, such as counterfactual CL (Roschewitz et al., 2025), DINO (Oquab et al., 2023), and hard-negative samples (Chen et al., 2024e; Wu et al., 2025b) to improve models' generalization ability and also multi-task contrastive learning Framework (Zavar et al., 2025b; Mendel et al., 2023b), and also local CL (Mendel et al., 2023b; Basak et al., 2023b) for capturing more fine-grained representation. However, the quality of the learned representations and the model's generalizability are highly dependent on the data augmentation strategy (Xiao et al., 2025a).

Then, adversarial learning (AL) deliberately introduces slight modifications to increase the difficulty of the training task (Lan et al., 2024a; Zhao et al., 2024), where the most common method is the Generative Adversarial Networks (GAN) framework (Yan et al., 2025b; Lee, 2023; Salimans et al., 2016). The GAN trained two components, one is the generator, which add noise on input data to generate adversarial samples, and the other one is the discriminator, which aims to discriminate a real data or a generated data (Li et al., 2024a, 2023d). This method significantly improves robustness in noisy scientific data, such as distinguishing chemically similar molecules with different biOA.ctivity (Lu et al., 2025), predicting proteins' binding-affinity changes upon mutation (Lan et al., 2024a). However, when AL provided with high-quality discrimination, it reduce the training speed (Xiao et al., 2025a) 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 MLLM 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 improves the alignment with clinical standards (Wang et al., 2025c,b; Lu et al., 2024c), while AutoProteinEngine (Liu et al., 2025e) demonstrates the use of literature retrieval to integrate protein engineer-

ing knowledge. Other systems employ real-time search engines to enrich responses with up-to-date information (Lu et al., 2024g; Ghafarollahi and Buehler, 2024; Lei et al., 2024b). Beyond direct retrieval, task-specific strategies have been explored, such as i-MedRAG (Xiong et al., 2024a), 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)** KG provides 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. (Aamer et al., 2025) combines LLMs with biomedical graphs to autonomously design research proposals and plan experiments. 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., 2024; Wei et al., 2023; Huang et al., 2023a), all without optimizing models' parameters. This approach has proven effective not only for scientific sequence data (Wang et al., 2023c) but also for multimodal data, including symbolic 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., 2024c; Nguyen and Grover, 2024).

Demonstration pairs play a central role in influencing ICL performance (Dong et al., 2024), 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 similar 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., 2023d). However, excessive demonstrations can also harm performance (Dong et al., 2024; Bertsch et al., 2025), 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 for improving multi-step reasoning ability (Oh et al., 2025; Shi et al., 2025).

**Planning** Frameworks such as ReAct (Yao et al., 2023), 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., 2022b) 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 (MarkTechPost, 2024) 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., 2024). Early systems like ChemCrow (Bran et al., 2023), BioDiscoveryAgent (Roohani et al., 2024), 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., 2025c) and AutoBA (Zhou et al., 2024a) 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. For utilizing laboratory hardware, a valuable resource is technical documentation (Boiko et al., 2023).

**Multi-step Reasoning** Chain-of-thought improves step-wise inference in scientific tasks (Chen et al., 2025b; Huang et al., 2024). Role-specialized agents with domain tools yield grounded results, for example, MT-Mol combines RDKit for chemical reasoning (Kim et al., 2025a) for more accurate molecules synthesis design. 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.

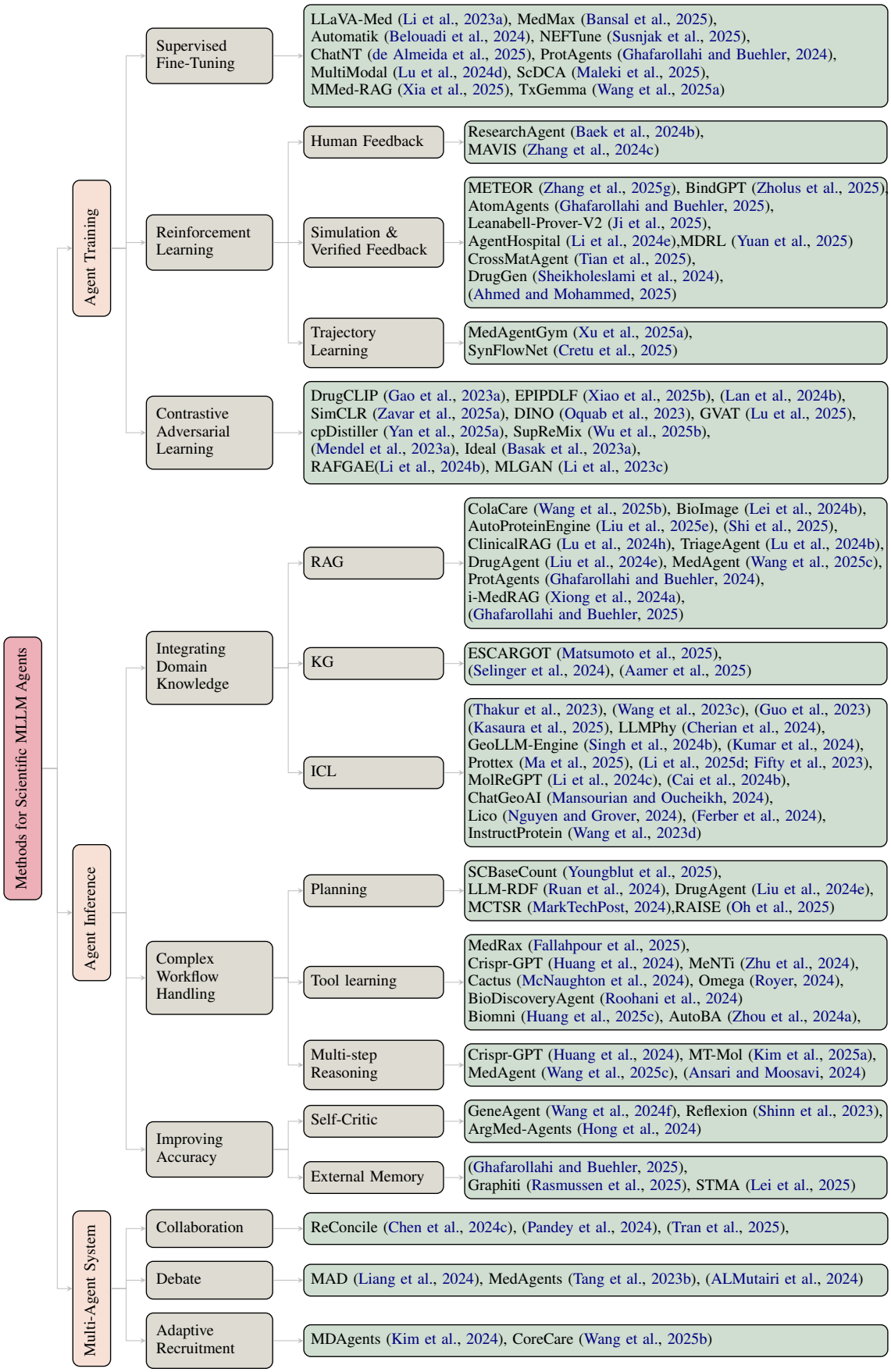


Figure 5: Taxonomy of Methods for Scientific MLLM Agents.

| 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). In each supplementary table, we 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 | | | | | Task Sources | Test Scale | Metrics |
|---|---|---|---|---|---|---|---|---|---|
| | | Seq. | Symb. | Vision | Graph | Other | | | |
| **MoleculeQA** (Lu et al., 2024f) | QA | ✓ | | | | | PubLib. | 62k | OA., Error |
| **PGxQA** (Keat et al., 2024) | QA | ✓ | | | | | PubLib., Man. | 110k | OA., Halluc., Man., Sim., Suc. |
| **PubMedQA** (Jin et al., 2019) | QA | ✓ | | | | ✓Tabular | Lit. | 1k | OA. |
| **MicroVQA** (Burgess et al., 2025) | QA | ✓ | | ✓ | | | Man. | 1k | OA. |
| **MedCalc-Bench** (Khandekar et al., 2024) | Medical Calculation | ✓ | | | | | PubLib. | 55 | OA. |
| **Bio-Benchmark** (Jiang et al., 2025) | Biological Knowledge | ✓ | | | | ✓Tabular | PubLib. | 1k | Halluc., Sim., Other |
| **BioLLMBench** (Sarwal et al., 2023) | Biological Knowledge | ✓ | ✓ | | | | Lit. | 36 | OA., Sim., Suc., Cost, Simplicity |
| **ProteinLMBench** (Shen et al., 2024) | Protein Understanding | ✓ | | | | | Lit., PubLib. | 944 | OA. |
| **LAB-Bench** (Laurent et al., 2024) | Biology Lab Research | ✓ | ✓ | ✓ | | ✓Tabular | Lit., Man., Syn. | 2.4k | OA. |
| **BioCoder** (Tang et al., 2023a) | Code Generation | ✓ | ✓ | | | | GitHub, Web. | 70k | OA. |
| **BioProBench** (Liu et al., 2025f) | Protocol Reasoning | ✓ | ✓ | | | ✓Tabular | PubLib., Syn. | 556k | Error, OA., Sim. |
| **scEval** (Liu et al., 2024f) | Single-cell Analysis | ✓ | | | | | Lit. | 22 | OA. |
| **SimBench** (Cao et al., 2021) | Single-cell Analysis | ✓ | | | | | Lit. | 8 | OA., Sim. |
| **DUD-E** (Mysinger et al., 2012) | Molecular Docking | ✓ | | | ✓ | ✓Tabular | PubLib., Syn. | 102 | OA. |
| **BaisBench** (Luo et al., 2025a) | Omics Discovery | ✓ | | | | | Lit., Syn. | 229 | OA. |

Table 7: Benchmarks on scientific MLLM agents (Biology)

| 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., 2025b) | Chemical Reasoning and Design | ✓ | ✓ | | ✓ | | Lit., Human. | 51 | OA., Novelty, Man. |
| **ChemLLMBench** (Guo et al., 2023) | Chemical Reasoning and Design | ✓ | | ✓ | | | PubLib. | 800 | OA., Man., Sim. |
| **MaSQA** (Zaki et al., 2023) | Materials Science Question Answering | ✓ | ✓ | | | | Lit., Man. | 650 | OA. |

Table 8: Benchmarks on scientific MLLM agents (Chemistry & Materials Science)

| 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., 2024b) | Multimodal Math Problem Solving | ✓ | ✓ | ✓ | | | PubLib. | 15.6k | OA., PSuc. |
| **MathChat** (Wu et al., 2023b) | Conversational Math Problem Solving | ✓ | ✓ | | | | PubLib., Syn. | 1.1k | OA., Sim., Error |
| **GeoEval** (Zhang et al., 2024a) | Geometry Problem Solving | ✓ | ✓ | ✓ | | | PubLib. | 5k | OA., PSuc. |
| **MathVista** (Lu et al., 2024e) | 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., 2023a) | 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., 2024b) | 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 (Ronchini et al., 2022)** | Multimodal Science QA | ✓ | | ✓ | | | PubLib., Man. | 21.2k | OA., Sim., Suc. |
| **MMSciCap+ MMSciQA (Li et al., 2025e)** | 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., 2024d)** | College Scientific Reasoning | ✓ | ✓ | | | | Lit., Man. | 972 | OA., Effective |
| **GPQA (Rein et al., 2024)** | Graduate-Level Expert QA | ✓ | | | | | Lit., Man. | 448 | OA., Objectivity |
| **SuperGPQA (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., 2024b)** | 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., 2018)** | Scientific Figure Visual QA | ✓ | | ✓ | | ✓Tabular | Lit., Man. | 15 | OA., Smoothness |
| **ArXivQA (Li et al., 2024f)** | 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., 2024a)** | Scientific Ability Evaluation | ✓ | | | | | PubLib., Lit., Man., Syn. | 18k | OA. |
| **MMEVALPRO (Huang et al., 2025a)** | Multimodal Benchmark Calibration | ✓ | ✓ | ✓ | | | Lit. | 6.4k | OA., Consistency |
| **ScienceAgentBench (Chen et al., 2025c)** | 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)