AI Agents for the Science of Science: A Survey of Tasks, Architectures, Evaluations, and Challenges

Yixuan Liu[*1], **Yicheng Zhang**[*1],
[1]Network Science Institute, Northeastern University
*Equal contributions.
Correspondence: liu.yixuan2@northeastern.edu

Abstract

The Science of Science (SciSci) examines how scientific knowledge is generated, evaluated, and transformed by utilizing large-scale scholarly and bibliometric data. As these data grow in scale and complexity, analysis has increasingly relied on statistical, network-based, machine learning methods, and is now seeing growing involvement of AI agents. This emerging class of such agents, ranging from multi-agent simulations of scientific behavior to tool-augmented systems for empirical analysis, is beginning to reshape how SciSci research is conducted. In this survey, we propose a task-centered taxonomy, distinguishing *agents as simulations*, which model citation, collaboration, and community dynamics, from *agents as tools*, which assist empirical analysis and scientific workflows. We review agent architectures, learning mechanisms, evaluation, and SciSci benchmarks, and examine open challenges related to reliability, data quality, and bias. Our survey aims to clarify the landscape of AI agents in SciSci and to support the development of reliable and scientifically useful AI systems for studying science and scientific communities.

1 Introduction

The Science of Science (SciSci) is an interdisciplinary research area that seeks to understand how scientific knowledge is produced, evaluated, and transformed over time, with particular attention across scientists, institutions, and scientific communities (Fortunato et al., 2018). By analyzing large-scale data on publications, citations, collaborations, and careers, SciSci seeks to uncover regularities and mechanisms underlying scientific progress, innovation, and inequality across individuals, teams, institutions, and fields (Fortunato et al., 2018; Wang et al., 2013; Uzzi et al., 2013; Liu et al., 2023; Wang and Barabási, 2021).

Addressing these questions has required an evolving methodological toolkit built around large-scale bibliometric data (Fortunato et al., 2018; Liu et al., 2023), with statistical and econometric models, causal inference, and network analysis used to study productivity, impact, innovation (Uzzi et al., 2013), and career dynamics (Wang et al., 2013; Petersen et al., 2014; Azoulay et al., 2011; Fortunato et al., 2018; Wu et al., 2019). More recently, machine learning and representation learning methods have enabled predictive modeling and pattern discovery by integrating scientific text, network structure, and temporal dynamics (Yan et al., 2011; Perozzi et al., 2014; Grover and Leskovec, 2016; Beltagy et al., 2019; Cohan et al., 2020). Beyond retrospective analysis, SciSci increasingly informs high-stakes contexts such as research evaluation and science policy, creating reflexive feedback between analysis and scientific behavior and placing heightened demands on robustness, interpretability, and multi-step reasoning (Hicks et al., 2015; Shao et al., 2025).

In parallel, recent advances have given rise to agent-based AI systems that embed language models within explicit control loops for planning (Yao et al., 2023b; Wang et al., 2023), tool use (Yao et al., 2023b), memory (Shinn et al., 2023; Hu et al., 2025), and iterative interaction with the external environment (Park et al., 2023), enabling multi-step, goal-directed behavior beyond single-shot generation. These capabilities position AI agents, particularly large language model (LLM) based agentic systems, as a well-suited computational framework for orchestrating complex, multi-step, and data-intensive scientific workflows (Masterman et al., 2024; Luo et al., 2025; Xi et al., 2025).

Important questions naturally arise: how can AI agents effectively support SciSci research, through which architectures and execution paradigms, and under what evaluation criteria? In this survey, we synthesize existing work on AI agents for SciSci while identifying key limitations, open challenges,

Findings of the Association for Computational Linguistics: ACL 2026, pages 36196–36211
July 2-7, 2026

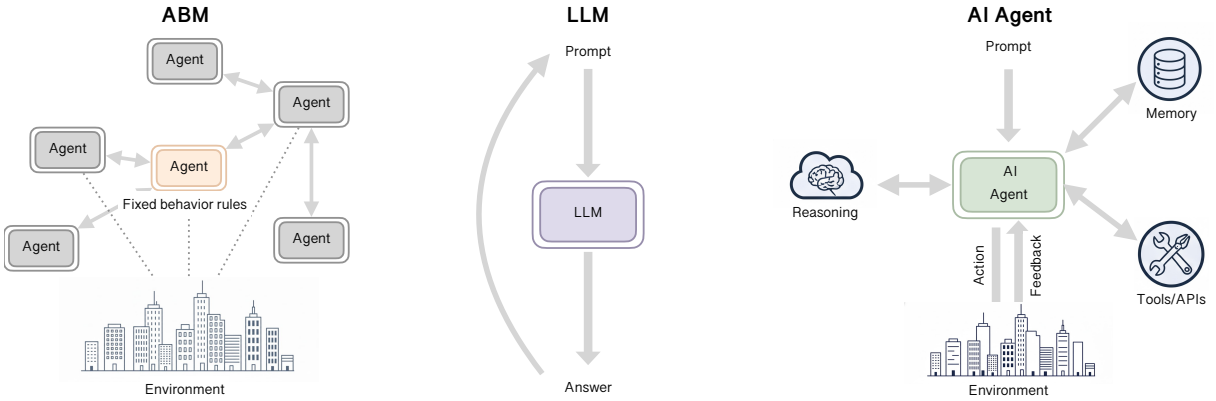

Figure 1: Comparison of three computational paradigms: (a) ABM with multiple agents following fixed behavior rules interacting within an environment, (b) LLM operating as a simple prompt-in, answer-out system with iterative generation, and (c) AI Agents combining an LLM core with reasoning, memory, and tools/APIs to take actions and receive feedback from the environment.

and opportunities for future research. We first analyze the core capabilities enabling AI agents in SciSci research in § 2. In § 3, we present a task-centered taxonomy, followed by a review of representative agent architectures and system paradigms (§ 4) and a summary of the underlying learning paradigms (§ 5). Finally, we examine evaluation metrics and benchmarks in § 6, and outline open challenges and implementation gaps for AI agents in SciSci research in § 7.

2 Agent Capabilities for SciSci Research

Among the diverse computational approaches to SciSci, a notable trajectory runs from simulation-based to language-based to agentic paradigms (Table 1, Figure 1). Early on, agent-based modeling (ABM) played a foundational role in SciSci by formalizing fixed micro-level behavioral rules about individual scientists and examining how their interactions give rise to macro-level patterns, like collaboration networks or citation dynamics (Gilbert, 1997; Watts and Gilbert, 2011; Chacko et al., 2025), directly supporting SciSci's goal of linking individual behavior to collective scientific outcomes.

Recent advances in AI agents mark a qualitative shift beyond traditional ABMs by enabling adaptive, data-driven decision-making and the orchestration of multi-step analytical workflows through rule-based planners, reinforcement learning agents, and LLM-based agentic systems (Wang et al., 2024a; Huang et al., 2024; Yao et al., 2023b; Shinn et al., 2023; Zhang et al., 2025). Consequently, AI agents can model micro-level human-like decision processes at scale, enabling more flexible analyses of macro-level scientific structures (Xie et al., 2024). These capabilities are especially well-suited to SciSci's distinctive demands: cross-disciplinary breadth spanning all of science, validation against established bibliometric regularities rather than domain-specific ground truth, and the reflexive challenge that agents studying science inevitably become part of the system they analyze (Fortunato et al., 2018). AI agents offer several key advantages over classical ABMs in SciSci including:

Reasoning and planning. Unlike static models responding to isolated queries, AI agents decompose complex tasks into manageable subtasks, execute them adaptively, and refine their approach based on intermediate results (Wang et al., 2024a). This autonomous orchestration capability, as in Figure 1 is particularly valuable for SciSci, where typical analyses rely on millions of papers (Lin et al., 2023; Fortunato et al., 2018) require chaining literature retrieval, data processing, and statistical modeling into coherent workflows. Agents achieve this through planning mechanisms such as Chain-of-Thought prompting (Wei et al., 2023), Tree-of-Thoughts exploration (Yao et al., 2023a), and ReAct reasoning-action loops (Yao et al., 2023b). These techniques enable handling SciSci questions that inherently span multiple temporal scales, disciplinary boundaries, and various analytical levels.

Tool use and knowledge integration. AI agents address fundamental LLM limitations: knowledge cutoffs, numerical imprecision, lack of live data via external sources and tool integration (Schick et al., 2023; Qu et al., 2025). Tool-augmented agents like SciSciGPT enable unified bibliomet-

Dimension	ABM	LLM	AI Agent
Mechanism	Rule-based, bottom-up emergence	Neural text generation	LLM + planning + tools + memory
Behavior Modeling	Predefined rules (e.g., preferential attachment)	Implicit patterns from training data	Adaptive, persona based, emergent dynamics
Data Integration	Manual parameter calibration	Training corpus; knowledge cutoff	Real-time tool access; RAG
Scalability	Computationally intensive	Efficient at scale	Orchestrates complex workflows
Autonomy	Predefined	Reactive	Proactive
SciSci Uses	Network formation, diffusion dynamics	Semantic analysis, classification	Simulation + workflow + hypothesis
Limitations	Rigid rules; calibration burden	Limited autonomy; hallucinations	Cost; reliability; complexity

Table 1: Comparison of related computational paradigms for SciSci research. ABMs established simulation-based SciSci with fixed behavioral rules (Epstein and Axtell, 1996; Bonabeau, 2002; Gilbert, 1997). LLMs enabled scalable text analysis but lack autonomy (Brown et al., 2020; Beltagy et al., 2019; Ji et al., 2023). AI Agents integrate LLM capabilities with planning, tool use, and adaptive behavior, bridging simulation and analysis (Wang et al., 2024a; Yao et al., 2023b; Park et al., 2023; Gao et al., 2024).

ric, algorithmic, and full-text analysis workflows over 134M papers (Shao et al., 2025), while PaperQA2 grounds claim verification in live literature through iterative retrieval and citation-aware reranking (Skarlinski et al., 2024). VirSci (Su et al., 2025) further demonstrates that, for ideation, integrating diverse tools: retrieval, network analysis, and semantic modeling, outperforms single capability pipelines.

Scalable semantic analysis. While traditional SciSci faces a tradeoff between qualitative depth (close reading of few texts) and quantitative breadth (bibliometrics over millions), LLM-based agents bridge this divide (Wei et al., 2025; Chen et al., 2025a). Agents comprehend scientific arguments, identify methodological choices, characterize rhetorical strategies, and synthesize thematic patterns across large corpora (Baek et al., 2025; Li et al., 2025). This enables content-aware SciSci research at scale: analyzing not just citation networks, but how ideas are framed, contested, and evolved across scientific discourse.

Human behavior simulation. Unlike traditional ABMs that rely on predefined rules, LLM-based agents can simulate nuanced scientist behavior. Trained on scientific literature, peer reviews, and academic discourse (Brown et al., 2020; Soldaini et al., 2024), LLMs internalize patterns of scientific reasoning and social interaction, enabling realistic simulation of scientist personas with associated biases, expertise, and dynamics (Park et al., 2023). Key enabling mechanisms include persona conditioning, memory-based behavioral consistency, multi-agent interaction protocols, and emergent system-level dynamics. Domain-specific systems such as AgentReview (Jin et al., 2024) and CiteAgent (Ji et al., 2025) have demonstrated these capabilities for simulating peer review and citation behavior in SciSci research.

Hypothesis generation and iterative refinement. AI agents can generate novel research hypotheses through multi-agent debate, self-critique, and iterative improvement (Lu et al., 2024; Ghafarollahi and Buehler, 2024). Frameworks like AI Co-Scientist (Gottweis et al., 2025) refine hypotheses through multi-agent tournament-style selection—a process that mirrors how scientific communities filter ideas through competitive evaluation (Fortunato et al., 2018). For SciSci, this capability serves dual purposes: agents can generate hypotheses about scientific phenomena, while the ideation process itself becomes an object of SciSci inquiry, revealing how AI systems internalize and reproduce patterns from scientific training data (Si et al., 2024).

3 Taxonomy

As above, many core SciSci challenges are inherently process-centric and interactive, making them well suited to AI agents. We propose a task-based taxonomy with two roles: **agents as simulations**, which model scientific processes to test mechanistic hypotheses; and **agents as tools**, which operationalize SciSci constructs—such as knowledge recombination, review bias, and consensus formation—to assist empirical SciSci analysis

as in Table 2.

3.1 Agents as Simulations

Modeling citation networks. Citation networks exhibit stable regularities such as power-law degree distributions and preferential attachment (Wang and Barabási, 2021; Fortunato et al., 2018). Ji et al. (2025) introduces CiteAgent, an LLM-based multi-agent framework that models authors as interacting agents and reproduces these empirical patterns, enabling controlled interventions on citation behavior. And reveals that different language models induce systematically different attachment dynamics. This demonstrates how agentic simulations can be used in SciSci to move beyond descriptive analysis toward mechanistic testing of citation formation.

Simulating collaboration dynamics. Scientific collaboration is the dominant mode of modern knowledge production (Wuchty et al., 2007), yet its underlying mechanisms are difficult to study observationally. Recent work leverages multi-agent simulations to model research communities and collaboration processes. Yu et al. (2025) propose ResearchTown, which represents the scientific ecosystem as an agent–data graph linking researcher agents and paper artifacts, enabling simulation of idea propagation and counterfactual collaboration scenarios. Complementary approaches model collaboration at the team level: VirSci (Su et al., 2025) demonstrates that structured multi-agent collaboration yields higher-novelty ideas than single-agent baselines, while Agent Laboratory and AgentRxiv (Schmidgall et al., 2025; Schmidgall and Moor, 2025) explicitly simulate hierarchical roles and cumulative knowledge building. Together, these agentic simulations provide a computational framework for testing SciSci theories of collaboration, team structure, and knowledge accumulation beyond static network analysis.

Simulating peer review dynamics. Peer review is both a quality control mechanism and a subject of SciSci inquiry (Fortunato et al., 2018; Liu et al., 2023). Jin et al. (2024) present AgentReview, an LLM-based framework simulating peer review with Reviewer, Author, and Area Chair agents, enabling controlled experiments on reviewer biases impossible with real data.

3.2 Agents as Tools

Each agent in this section is included on a specific criterion: it must operationalize an established SciSci theoretical construct: knowledge recombination, peer review norms, or evidence accumulation — such that its design is informed by SciSci theory and its outputs can be evaluated against known scientific regularities (Fortunato et al., 2018). This distinguishes the systems here from general research automation: rather than merely accelerating scientific tasks, they make scientific processes computationally observable and measurable, serving simultaneously as instruments *for* SciSci research and objects *of* SciSci inquiry. This grounding places them at the intersection of SciSci theory and scientific practice (Chen et al., 2025b).

Automated review generation. Complementing simulation-based studies, other agents focus on *generating* useful feedback as a practical tool. Peer review constitutes the primary gatekeeping mechanism through which scientific knowledge is validated, yet the implicit norms governing review quality have remained difficult to formalize (Fortunato et al., 2018). MARG (D'Arcy et al., 2024) deploys a leader–worker–expert hierarchy to produce actionable comments, reducing generic feedback from 60% to 29%. REMOR (Taechoyotin and Acuna, 2025) uses reinforcement learning from human preferences to align generated reviews with expert judgment. By encoding reviewer behavior as learnable objectives, these systems make the implicit norms of scientific gatekeeping explicit and measurable — revealing that review quality is both decomposable and improvable.

Research idea generation. Understanding how novel scientific ideas emerge is a longstanding SciSci challenge (Wang and Barabási, 2021; Uzzi et al., 2013). A key insight from SciSci is that high-impact discoveries often arise from *atypical combinations* of otherwise conventional knowledge (Uzzi et al., 2013); recent systems operationalize this knowledge recombination theory by mining implicit connections across corpora to generate novel research ideas, enabling the study of ideation as a computational process (Zheng et al., 2025). One line of work focuses on automated generation pipelines. Systems like SciMON (Wang et al., 2024b) retrieve related papers and iteratively refine hypotheses using novelty boosting. ResearchAgent (Baek et al., 2025) extends this with multi-agent feedback from specialized ReviewingAgents, while SciAgents (Ghafarollahi and Buehler, 2024) incorporates ontological reasoning over knowledge graphs for cross-domain hypothesis discovery.

A complementary line emphasizes human-AI co-ideation and empirical evaluation. Scideator

Tasks	Core SciSci Question	Representative Systems
Agents as Simulations: modeling scientific process		
Citation network modeling	How do citations emerge? Patterns?	CiteAgent (Ji et al., 2025)
Collaboration dynamics	How do scientists collaborate and research communities form?	ResearchTown (Yu et al., 2025), VirSci (Su et al., 2025), Agent Laboratory (Schmidgall et al., 2025), AgentRxiv (Schmidgall and Moor, 2025)
Peer review dynamics	How does peer review function?	AgentReview (Jin et al., 2024)
Agents as Tools: assisting SciSci analysis		
Automated review generation	How to generate useful reviewer feedback?	MARG (D'Arcy et al., 2024), REMOR (Taechoyotin and Acuna, 2025)
Research idea generation	How does knowledge recombination drive novelty in science?	SciMON (Wang et al., 2024b), ResearchAgent (Baek et al., 2025), SciAgents (Ghafarollahi and Buehler, 2024), Scideator (Radensky et al., 2025), CoQuest (Liu et al., 2024)
Scientific claim verification	How does evidence accumulate and scientific consensus form?	SciFact (Wadden et al., 2020), PaperQA2 (Skarlinski et al., 2024)
SciSci workflow automation	How can we automate SciSci research?	SciSciGPT (Shao et al., 2025), DataParasite (Sun, 2026)

Table 2: Task-based taxonomy of AI agents for SciSci. *Simulations* model scientific processes to test mechanistic hypotheses; *Tools* operationalize SciSci theories and constructs—such as knowledge recombination, peer review process, and consensus formation—to assist empirical analysis.

(Radensky et al., 2025) operationalizes combinatorial innovation theory by decomposing papers into facets (purpose, mechanism, evaluation) and letting users interactively recombine them. CoQuest (Liu et al., 2024) studies human-AI co-creation through branching exploration of research questions. Most ambitiously, Si et al. (2024) conduct the first large-scale blind study with over 100 NLP researchers, finding that LLM-generated ideas are rated significantly more novel but less feasible than human ideas, with experts unable to reliably distinguish between them. Together, these systems make ideation observable and measurable, turning knowledge recombination from a theoretical construct into an empirical SciSci object.

Scientific claim verification. The accumulation and verification of scientific claims is fundamental to knowledge production (Fortunato et al., 2018). From a SciSci perspective, how evidence accumulates across literature and how scientific consensus forms are central epistemological questions (Fortunato et al., 2018; Wang and Barabási, 2021). Verification agents operationalize these questions through retrieval-augmented pipelines that can be studied as models of evidence aggregation. SciFact (Wadden et al., 2020) established the task of classifying claims as Supported, Refuted, or Not Enough Info given evidence abstracts. Skarlinski et al. (2024) extend this with PaperQA2, an agentic system achieving 85.2% precision on the LitQA2 benchmark, surpassing the 73.8% achieved by human domain experts. The system employs iterative query refinement, citation-aware ranking, and self-evaluation loops. The gap between agent and expert performance itself constitutes a SciSci finding about the structure of domain knowledge and the limits of evidence synthesis at scale.

SciSci workflow automation. Beyond task-specific agents, recent work develops integrated AI systems that automate end-to-end SciSci workflows over large-scale data. Shao et al. (2025) introduces SciSciGPT, a multi-agent research assistant that orchestrates literature retrieval, database querying over SciSciNet (134M papers), statistical analysis, and self-evaluation through a central manager agent. Designed explicitly for SciSci tasks: such as collaboration network analysis and atypical combination discovery, the system demonstrates how agentic architectures can support scalable, reproducible SciSci research. Complementing such workflow-level agents, tools like DataParasite (Sun, 2026) focus on automatically collecting and structuring large-scale Science of Science data, further lowering the barrier to empirical and agent-driven SciSci analysis.

The two paradigms are complementary: **simulations** test mechanistic hypotheses about how science works, while **tools** operationalize SciSci

theories—such as knowledge recombination, review norms, and evidence aggregation—to augment researchers' analytical capacity. Crucially, both paradigms reflect a bidirectional relationship: SciSci theories guide agent design across scientific disciplines, while agent-generated outputs—simulated citation networks, AI-produced reviews, computationally discovered ideas—create new empirical objects for SciSci inquiry (Chen et al., 2025b).

4 Architecture

SciSci agent systems exhibit recurring architectural patterns in how they organize control, tool use, memory, and coordination. These choices are consequential because SciSci workflows interleave literature retrieval, data analysis, and writing, where intermediate artifacts matter as much as final outputs.

Single-agent architectures. Single-agent designs suit SciSci tasks that unfold as sequential, self-contained workflows. The AI Scientist (Lu et al., 2024) automates the full research pipeline—ideation, experimentation, and writing—using a single LM controller. SciMON (Wang et al., 2024b) iteratively refines hypotheses by retrieving related work and penalizing similarity to prior ideas. REMOR (Taechoyotin and Acuna, 2025) trains a single reasoning model with reinforcement learning for review generation. These examples suggest that when tasks have well-defined boundaries and rely on iterative agent-knowledge interaction, single-agent architectures offer sufficient expressiveness without coordination overhead.

Multi-agent architectures. Multi-agent architectures distribute responsibilities across interacting agents. VirSci organizes agents into propose-critique-revise cycles for idea refinement (Su et al., 2025). ResearchTown models a research community as an agent–data graph, implementing reading, writing, and reviewing as message-passing interactions (Yu et al., 2025). CiteAgent couples agent decisions to a networked environment, enabling evaluation of micro-level behaviors through their effects on citation dynamics (Ji et al., 2025). The central architectural question is how information access, responsibilities, and feedback are structured.

Many systems make roles explicit to constrain tool access, memory scope, and interaction patterns. Common roles include planning, literature retrieval, data analysis, writing, and critical review. Table 3 summarizes a role taxonomy; Table 5 maps these roles to representative systems, showing how SciSci tasks require different role configurations.

Hybrid architectures. Most practical SciSci systems adopt hybrid architectures that combine language models with retrieval-augmented generation (RAG), tools, and persistent memory (Lewis et al., 2020; Liu et al., 2025a). PaperQA2 integrates staged retrieval with citation-aware ranking (Skarlinski et al., 2024); Scideator combines retrieval with facet-based ideation (Radensky et al., 2025). Memory modules extend agent capabilities beyond fixed context windows, as in ResearchTown's persistent agent–paper relations (Yu et al., 2025). An idealized workflow is provided in Appendix A.3.

5 Learning Paradigms

A notable pattern emerges: most surveyed systems employ pre-trained foundation models without fine-tuning, leveraging inference-time mechanisms including retrieval augmentation, tool integration, and structured prompting than specialized training.

5.1 Supervised/Instruction-Based Learning

Pre-training foundations. Most SciSci agents leverage general-purpose LLMs without domain-specific pre-training, relying on knowledge embedded during foundation model pre-training due to their strong zero-shot performances, supplemented by retrieval at inference time (Qi et al., 2023).

Supervised fine-tuning. Fine-tuning remains underexplored in SciSci agents. REMOR (Taechoyotin and Acuna, 2025) applies LoRA-based SFT on $\sim$1,700 peer review samples with synthetic reasoning traces before reinforcement learning. BioAgents (Mehandru et al., 2025) fine-tunes smaller models on domain documentation to achieve expert-comparable results. SFT can effectively encode domain conventions and scientific patterns, and remains limited by data availability.

5.2 Preference Optimization

Reward-based methods. Preference optimization remains underexplored in SciSci. REMOR (Taechoyotin and Acuna, 2025) applies Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a critic-free reinforcement learning variant, with algorithmically derived rewards across eight review dimensions, achieving $> 2\times$ human baseline scores. Such approaches are suitable for domains where expert feedback is costly, but quality metrics are formalized.

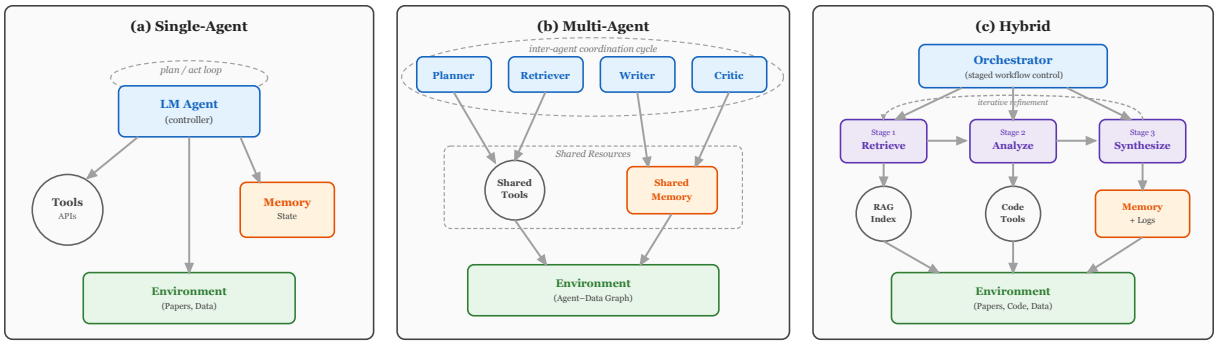

Figure 2: Architectural patterns in SciSci agent systems. **(a) Single-Agent**: A single LM controller iterates through reasoning, tool use, and memory access. **(b) Multi-Agent**: Role-specialized agents coordinate through shared tools and memory. **(c) Hybrid**: An orchestrator manages staged workflows with retrieval, analysis, and synthesis modules. Solid arrows are data and control flow; dashed lines denote coordination mechanisms and iterative refinement loops.

Role Category	Canonical Functions	Typical Responsibilities	Representative Instantiations
Planner and orchestrator	Task decomposition, workflow coordination, resource allocation	Goal parsing, subtask generation, agent scheduling, progress monitoring	Supervisor (Gottweis et al., 2025), Orchestrator (Liu et al., 2025b), Area Chair (Jin et al., 2024), Leader (D'Arcy et al., 2024)
Literature specialist	Information retrieval, citation management, knowledge grounding	Semantic search, query expansion, document ranking, citation verification	Paper Search Specialist (Liu et al., 2025b), Retriever (Yu et al., 2025), RAG modules (Skarlinski et al., 2024)
Domain expert or analyst	Data processing, method execution, domain-specific reasoning	Statistical analysis, code execution, experiment design, domain inference	Experiments Agent (D'Arcy et al., 2024), Scientist 1 & 2 (Ghafarollahi and Buehler, 2024), Data Analyst (Schmidgall et al., 2025)
Writer or synthesizer	Content generation, narrative construction, document assembly	Drafting, summarization, cross-source integration, format compliance	Academic Survey Writer (Liu et al., 2025b), Writer (Yu et al., 2025), Author (Jin et al., 2024)
Critic or evaluator	Quality assessment, feedback generation, iterative refinement	Novelty/soundness evaluation, weakness identification, score assignment	Critic (Ghafarollahi and Buehler, 2024), Reviewer (Jin et al., 2024), Quality Evaluator (Liu et al., 2025b), Impact Expert (D'Arcy et al., 2024)

Table 3: Common agent roles in SciSci systems. Each role category encompasses a cluster of related functions that recur across systems, though specific implementations vary in granularity and capability scope.

Reward-free methods. These methods appear in specific scientific domains for drug design (Cheng et al., 2024), remaining absent from general SciSci. Given Direct Preference Optimization (DPO)'s success in summarization (Rafailov et al., 2024) and KTO's ability to learn from unpaired binary feedback (Ethayarajh et al., 2024), they offer potential for literature synthesis and peer review tasks where quality signals are abundant but paired preferences are scarce.

5.3 Inference-Time Adaptation

In-context learning. Role-based prompting dominates SciSci agents as we summarized in Table 3. AgentReview assigns reviewer personas with attributes for commitment, intention, and knowledgeability (Jin et al., 2024); MARG deploys leader-worker-expert hierarchies (D'Arcy et al., 2024); VirSci coordinates up to eight specialized agents through structured collaboration protocols (Su et al., 2025).

Retrieval augmentation. RAG architectures are widely adopted. PaperQA2 uses Reranking and Contextual Summarization (Skarlinski et al., 2024); SciMON employs triple retrieval (semantic, graph, citation) (Wang et al., 2024b); SciSciGPT uses vector search with metadata filtering (Shao et al., 2025); ResearchAgent and SciAgents integrate entity-centric stores (Baek et al., 2025; Ghafarollahi and Buehler, 2024). These retrieval mechanisms are crucial for SciSci tasks, enabling agents to ground analysis while continuously evolving scholarly databases and maintain factual accuracy.

Reflection and self-critique. Multi-agent cri-

tique loops enable iterative refinement without training. ResearchAgent's ReviewingAgents evaluate hypotheses on validity, novelty, and significance (Baek et al., 2025); Agent Laboratory implements hierarchical academic roles with human-in-the-loop validation (Schmidgall et al., 2025); SciMON applies novelty-boosting penalties during iterative generation (Wang et al., 2024b). SciSciGPT employs a prompting technique using structured XML tags and iterative feedback from an evaluation agent (Shao et al., 2025).

This landscape shows that SciSci agents have made remarkable progress through inference-time techniques, but have largely overlooked preference optimization—an underexplored frontier with substantial room for advancement.

6 Evaluation and Benchmarks

Evaluating AI agents for SciSci presents fundamental challenges: simulation agents require theory alignment validation, tool agents demand task-specific performance metrics.

6.1 Evaluation Approaches

Human Evaluation. Protocols differ substantially across systems. Idea generation systems typically employ expert panels: ResearchAgent uses 10 domain experts across 15 criteria ($\rho = 0.83$) (Baek et al., 2025); SciMON recruits 6 NLP experts (Wang et al., 2024b); CoQuest conducts think-aloud protocols with 20 doctoral students (Liu et al., 2024); and Scideator adopts the validated Creativity Support Index ($n = 22$) (Radensky et al., 2025). Workflow automation systems like SciSciGPT use smaller pilot studies ($n = 3$) with expert interviews (Shao et al., 2025). Notably, the AI Co-Scientist extends beyond ratings to wet-lab validation, confirming tumor inhibition for AI-proposed drug candidates (Gottweis et al., 2025). Simulation-focused systems such as CiteAgent, VirSci, and ResearchTown largely forgo human evaluation, instead validating against established bibliometric theories (Ji et al., 2025; Su et al., 2025; Yu et al., 2025).

Automatic Evaluation. Four dominant approaches have emerged: (1) LLM-as-judge methods rating outputs on novelty and feasibility; (2) embedding-based similarity using SciBERT or SentenceBERT; (3) SciSci-grounded metrics operationalizing bibliometric constructs, such as VirSci's composite novelty score (Su et al., 2025; Uzzi et al., 2013) and CiteAgent's validation against citation network laws (Ji et al., 2025); and (4) multi-dimensional frameworks like REMOR's 8-aspect Human-aligned Peer Review Reward (Taechoyotin and Acuna, 2025) and SciSciGPT's three-level self-evaluation (Shao et al., 2025). However, automatic metrics do not reliably approximate human judgment: Agent Laboratory finds LLM reviewers inflate scores by 60% compared to human raters, while ResearchTown reports only r=0.49 correlation between LLM and human assessments (Schmidgall et al., 2025; Yu et al., 2025).

6.2 Benchmarks and Datasets

Task-specific benchmarks. Benchmark proliferation over standardization characterizes the field: most systems introduce task-specific benchmarks. Idea generation systems contribute corpora ranging from 67 facet-annotated examples (Radensky et al., 2025) to 85K papers with novelty frameworks (Baek et al., 2025; Su et al., 2025). Peer review systems provide 16K–53K annotated reviews for dynamics analysis (Jin et al., 2024; Taechoyotin and Acuna, 2025). Claim verification benchmarks include SciFact (1.4K claims with rationales) (Wadden et al., 2020) and LitQA2 (248 questions including unanswerable items) (Skarlinski et al., 2024). Citation network studies leverage domain-specific collections such as LLM-Agent (165 papers, 2021–2023) (Ji et al., 2025). Cross-system benchmarks have also emerged: ScienceAgentBench for data-driven discovery (102 tasks, 32–34% agent completion) (Chen et al., 2024), MLE-Bench for end-to-end ML research (Chan et al., 2024), ResearchBench for hypothesis generation (2024 papers only, contamination-resistant) (Liu et al., 2025c), and ResearchTown's ResearchBench with 1.2K paper-reconstruction tasks (Yu et al., 2025).

Shared Corpora and emerging standards. Large-scale scientific corpora serve as foundational resources across systems: including S2ORC (Lo et al., 2020), OpenAlex (Priem et al., 2022), SciSciNet (Lin et al., 2023), AMiner (Tang et al., 2008), and Open Academic Graph (Zhang et al., 2019). Citation network studies leverage Cora, CiteSeer, and OGB-Citation2 (Ji et al., 2025), whose well-documented structural regularities, including scale-free degree distributions (Lu et al., 2025) make them natural benchmarks for validating whether agent-simulated citation dynamics reproduce known empirical regularities. Yet no unified benchmark currently reached, these shared resources provide a foundation for cross-system com-

Category	Major benchmarks and datasets
Idea Generation	ResearchAgent corpus (50K); VirSci (85K); SciMON CLBD; Scideator facets
Peer Review	AgentReview (53K); REMOR PeerRT+HPRR (16K); ResearchBench (1.2K)
Claim/QA	SciFact (1.4K claims); LitQA2 (248); PubMedQA (273K)
Automation	ScienceAgentBench (102); MLE-Bench (75); ResearchBench (1.4K)
Citation	Cora; CiteSeer; OGB-Citation2 (2.9M); SciSciNet, LLM-Agent
Corpora	S2ORC (81M); OpenAlex (250M); Semantic Scholar; SciSciNet(-V2); AMiner

Table 4: Benchmarks and datasets by selected task category. Numbers indicate scale (papers, tasks, or items).

parison. Bridging task-specific evaluations with shared benchmark infrastructure remains essential for systematic progress.

7 Opportunities and Challenges

AI agents offer SciSci a methodological shift: the ability to *intervene* in scientific processes, not merely observe them. Simulations enable counterfactual experiments on collaboration, citation, and peer review dynamics, while tool-augmented agents scale semantic analysis across millions of papers. These capabilities distinguish SciSci agents from those in coding or e-commerce, where agents are exclusively task-completion tools. In SciSci, agents serve a dual role as both instruments and objects of inquiry—an agent simulating citation dynamics produces behavior that is itself a SciSci datum (Ji et al., 2025). This reflexive structure, combined with evaluation against established scientific regularities rather than solely task-completion metrics (Uzzi et al., 2013; Fortunato et al., 2018), creates an epistemically grounded paradigm with few parallels in other domains.

Yet fundamental challenges remain. Progress has largely relied on inference-time techniques, while preference optimization for aligning agent behavior with scientific norms requires further development. Current systems show weak alignment between automated and human evaluation (Schmidgall et al., 2025), and hallucinations undermine reproducibility. More critically, agents trained on existing corpora risk amplifying the very biases SciSci seeks to study: LLMs exhibit more pronounced citation bias toward highly cited work than human authors (Algaba et al., 2025), while AI tool adoption expands individual productivity but contracts collective scientific focus (Hao et al., 2025). Addressing these issues requires more than improved benchmarks; it demands rethinking how AI agents are trained, aligned, and validated when their outputs increasingly become objects of SciSci inquiry.

Limitations

Scope and inclusion criteria. Our survey focuses on systems explicitly designed for SciSci research, while recognizing that SciSci, AI-for-Science, and scientific AI assistants form interconnected but distinct perspectives. We selectively include adjacent systems (e.g., AI Scientist) when they address core SciSci questions like how novel ideas emerge, how collaboration affects innovation, or how to automate its workflows. This reflects a deliberate focus on agents that study or operationalize scientific processes rather than merely perform scientific tasks.

Between agents and non-agent borderlines. The term *AI agent* is used inconsistently across communities. Many systems labeled as agents are better described as tool-augmented LLM pipelines, while some multi-stage systems without the label exhibit agentic properties (e.g., closed-loop planning, persistent memory, or environment interaction). Rather than adjudicate these disputes, we adopt a functional stance: systems are organized by their role in SciSci research (simulation vs. tool) and evaluated on capability primitives (planning, tool use, memory, environment feedback). Some inclusions will inevitably be contested; we treat this ambiguity as a feature of an emerging field rather than a flaw of our taxonomy.

Comparability and reproducibility. Direct comparison across agent systems remains challenging due to heterogeneity in task definitions, data access, tool stacks, and evaluation protocols. Moreover, many systems depend on closed-source LLMs, undisclosed prompts, or evolving APIs, making it difficult to isolate architectural contributions. Consequently, our survey prioritizes recurring capabilities, architectural patterns, and evaluation practices rather than definitive cross-system rankings.

Acknowledgments

We thank all the reviewers and ACs for their valuable comments. Generative AI tools were used solely for language refinement and grammatical correction.

References

Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025. Large language models reflect human citation patterns with a heightened citation bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6844–6879, Albuquerque, New Mexico. Association for Computational Linguistics.

Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso. 2011. Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, 42(3):527–554.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. *arXiv preprint*. ArXiv:2404.07738 [cs].

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv preprint*. ArXiv:1903.10676 [cs].

Eric Bonabeau. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl_3):7280–7287.

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

George Chacko, Minhyuk Park, Vikram Ramavarapu, Ananth Grama, Pablo Robles-Granda, and Tandy Warnow. 2025. An Agent-based Model of Citation Behavior. *arXiv preprint*. ArXiv:2503.06579 [cs].

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. 2024. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering.

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 [cs].

Renqi Chen, Haoyang Su, Shixiang Tang, Zhenfei Yin, Qi Wu, Hui Li, Ye Sun, Nanqing Dong, Wanli Ouyang, and Philip Torr. 2025b. AI-Driven Automation Can Become the Foundation of Next-Era Science of Science Research. *arXiv preprint*. ArXiv:2505.12039 [cs].

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. 2024. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery.

Xiwei Cheng, Xiangxin Zhou, Yuwei Yang, Yu Bao, and Quanquan Gu. 2024. Decomposed Direct Preference Optimization for Structure-Based Drug Design. *arXiv preprint*. ArXiv:2407.13981 [q-bio].

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. *arXiv preprint*. ArXiv:2004.07180 [cs].

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. MARG: Multi-Agent Review Generation for Scientific Papers. *arXiv preprint*. ArXiv:2401.04259 [cs].

Joshua M. Epstein and Robert L. Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Complex Adaptive Systems. MIT Press, Cambridge, MA, USA.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. *arXiv preprint*. ArXiv:2402.01306 [cs].

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science*, 359(6379):eaao0185.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259.

Alireza Ghafarollahi and Markus J. Buehler. 2024. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint*. ArXiv:2409.05556 [cs].

N. Gilbert. 1997. A Simulation of the Structure of Academic Science. *Sociological Research Online*, 2(2):91–105.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2025. Towards an AI co-scientist. *arXiv preprint*. ArXiv:2502.18864 [cs].

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA. Association for Computing Machinery.

Qianyue Hao, Fengli Xu, Yong Li, and James Evans. 2025. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 649:1237–1243.

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. 2015. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548):429–431.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2025. Memory in the Age of AI Agents. *arXiv preprint*. ArXiv:2512.13564 [cs].

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint*. ArXiv:2402.02716 [cs].

Jiarui Ji, Runlin Lei, Xuchen Pan, Zhewei Wei, Hao Sun, Yankai Lin, Xu Chen, Yongzheng Yang, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2025. Leveraging LLM-based agents for social science research: insights from citation network simulations. *arXiv preprint*. ArXiv:2511.03758 [physics].

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring Peer Review Dynamics with LLM Agents. *arXiv preprint*. ArXiv:2406.12708 [cs].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025. SciLitLLM: How to Adapt LLMs for Scientific Literature Understanding. *arXiv preprint*. ArXiv:2408.15545 [cs].

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. 2023. SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315.

Chengwei Liu, Chong Wang, Jiayue Cao, Jingquan Ge, Kun Wang, Lvye Zhang, Ming-Ming Cheng, Penghai Zhao, Tianlin Li, Xiaojun Jia, Xiang Li, Xinfeng Li, Yang Liu, Yebo Feng, Yihao Huang, Yijia Xu, Yuqiang Sun, Zhenhong Zhou, and Zhengzi Xu. 2025a. A vision for auto research with LLM agents. *Preprint*, arXiv:2504.18765.

Lu Liu, Benjamin F. Jones, Brian Uzzi, and Dashun Wang. 2023. Data, measurement and empirical methods in the science of science. *Nature Human Behaviour*, 7(7):1046–1058.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. CoQuest: Exploring Research Question Co-Creation with an LLM-based Agent. *arXiv preprint*. ArXiv:2310.06155 [cs].

Yixin Liu, Yonghui Wu, Denghui Zhang, and Lichao Sun. 2025b. Agentic autosurvey: Let LLMs survey LLMs. *Preprint*, arXiv:2509.18661.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025c. ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint*. ArXiv:2408.06292 [cs].

Jianglin Lu, Yixuan Liu, Yitian Zhang, and Yun Fu. 2025. Scale-Free Graph-Language Models. *arXiv preprint*. ArXiv:2502.15189 [cs].

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, and 7 others.

2025. Large Language Model Agent: A Survey on Methodology, Applications and Challenges. *arXiv preprint*. ArXiv:2503.21460 [cs].

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

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey. *arXiv preprint*. ArXiv:2404.11584 [cs].

Nikita Mehandru, Amanda K. Hall, Olesya Melnichenko, Yulia Dubinina, Daniel Tsirulnikov, David Bamman, Ahmed Alaa, Scott Saponas, and Venkat S. Malladi. 2025. BioAgents: Bridging the gap in bioinformatics analysis with multi-agent systems. *Scientific Reports*, 15(1):39036. Publisher: Nature Publishing Group.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint*. ArXiv:2304.03442 [cs].

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ArXiv:1403.6652 [cs].

Alexander Michael Petersen, Santo Fortunato, Raj K. Pan, Kimmo Kaski, Orion Penner, Armando Rungi, Massimo Riccaboni, H. Eugene Stanley, and Fabio Pammolli. 2014. Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences*, 111(43):15316–15321.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint*. ArXiv:2205.01833 [cs].

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large Language Models are Zero Shot Hypothesis Proposers. *arXiv preprint*. ArXiv:2311.05965 [cs].

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool Learning with Large Language Models: A Survey. *Frontiers of Computer Science*, 19(8):198343. ArXiv:2405.17935 [cs].

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2025. Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination. *arXiv preprint*. ArXiv:2409.14634 [cs].

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 [cs].

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint*. ArXiv:2302.04761 [cs].

Samuel Schmidgall and Michael Moor. 2025. AgentRxiv: Towards Collaborative Autonomous Research. *arXiv preprint*. ArXiv:2503.18102 [cs].

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 [cs].

Erzhuo Shao, Yifang Wang, Yifan Qian, Zhenyu Pan, Han Liu, and Dashun Wang. 2025. SciSciGPT: advancing human–AI collaboration in the science of science. *Nature Computational Science*, pages 1–15.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint*. ArXiv:2402.03300 [cs].

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 [cs].

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 [cs].

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint*. ArXiv:2409.13740 [cs].

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria. Association for Computational Linguistics.

Mengyi Sun. 2026. DataParasite enables scalable and repurposable online data curation. *arXiv preprint*. ArXiv:2601.02578 [cs].

Pawin Taechoyotin and Daniel Acuna. 2025. REMOR: Automated Peer Review Generation with LLM Reasoning and Multi-Objective Reinforcement Learning. *arXiv preprint*. ArXiv:2505.11718 [cs].

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 990–998, New York, NY, USA. Association for Computing Machinery.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical Combinations and Scientific Impact. *Science*, 342(6157):468–472.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Dashun Wang and Albert-László Barabási. 2021. *The Science of Science*. Cambridge University Press, Cambridge.

Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying Long-Term Scientific Impact. *Science*, 342(6154):127–132.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv preprint*. ArXiv:2305.16291 [cs].

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024a. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6):186345. ArXiv:2308.11432 [cs].

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)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Watts and Nigel Gilbert. 2011. Does cumulative advantage affect collective learning in science? An agent-based simulation. *Scientometrics*, 89(1):437–463.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, and 8 others. 2025. From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery. *arXiv preprint*. ArXiv:2508.14111 [cs].

Lingfei Wu, Dashun Wang, and James A. Evans. 2019. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382.

Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. 2007. The increasing dominance of teams in production of knowledge. *Science (New York, N.Y.)*, 316(5827):1036–1039.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 9 others. 2025. The rise and potential of large language model based agents: a survey. *Science China Information Sciences*, 68(2):121101.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? *arXiv preprint*. ArXiv:2402.04559 [cs].

Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1247–1252, New York, NY, USA. Association for Computing Machinery.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint*. ArXiv:2305.10601 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b.

ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint*. ArXiv:2210.03629 [cs].

Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. 2025. ResearchTown: Simulator of Human Research Community. *arXiv preprint*. ArXiv:2412.17767 [cs].

Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, and Kuansan Wang. 2019. OAG: Toward Linking Large-scale Heterogeneous Entity Graphs. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2585–2595, New York, NY, USA. Association for Computing Machinery.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Francisco Piedrahita-Velez, Yue Liao, Hongru Wang, and 6 others. 2025. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey. *arXiv preprint*. ArXiv:2509.02547 [cs].

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

A Appendix

A.1 Agent Roles in SciSci by Tasks

Table 5 presents a task-oriented taxonomy distinguishing two roles: *Agents as Simulations* model scientific processes to test mechanistic hypotheses, while *Agents as Tools* augment empirical analysis and workflow automation.

A.2 Architecture Classification of SciSci Agents

Table 6 provides a comprehensive classification of SciSci agent systems discussed in this survey according to their architectural patterns: single-agent, multi-agent, or hybrid.

A.3 Idealized Workflow for SciSci Agents

Figure 3 provides an idealized view of a hybrid SciSci agent workflow. The intent is to illustrate how retrieval, tools, memory, and (optionally) role separation can be composed into a staged pipeline, rather than to prescribe a single system design.

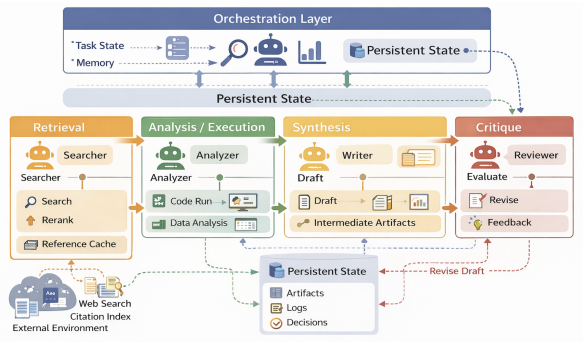

Figure 3: An idealized hybrid workflow for SciSci agent systems. A top-level orchestrator stages retrieval, analysis or execution, synthesis, and critique while maintaining persistent state such as evidence logs and intermediate artifacts. The diagram abstracts common design patterns observed across existing SciSci systems, illustrating how language models, tools, memory, and (optionally) role specialization can be composed into a coherent pipeline. It is intended as a conceptual summary rather than a prescriptive system architecture.

Tasks	Planner	Lit. Specialist	Analyst	Writer	Critic	Representative Systems
Agents as Simulations						
Citation network modeling	○	●	●	●	○	CiteAgent (Ji et al., 2025)
Collaboration dynamics	●	●	◐	●	●	ResearchTown (Yu et al., 2025), VirSci (Su et al., 2025), Agent Laboratory (Schmidgall et al., 2025), AgentRxiv (Schmidgall and Moor, 2025)
Peer review dynamics	●	○	○	●	●	AgentReview (Jin et al., 2024)
Agents as Tools						
Automated review generation	◐	○	◐	●	●	MARG (D'Arcy et al., 2024), REMOR (Taechoyotin and Acuna, 2025)
Research idea generation	◐	●	◐	●	●	SciMON (Wang et al., 2024b), ResearchAgent (Baek et al., 2025), SciAgents (Ghafarollahi and Buehler, 2024), Scideator (Radensky et al., 2025), CoQuest (Liu et al., 2024)
Scientific claim verification	◐	●	◐	◐	●	SciFact (Wadden et al., 2020), PaperQA2 (Skarlinski et al., 2024)
SciSci workflow automation	●	●	●	◐	●	SciSciGPT (Shao et al., 2025), DataParasite (Sun, 2026)

Legend: ● Explicitly instantiated; ◐ Partially presented; ○ Not present.

Table 5: Task-based taxonomy of AI agents for SciSci in selected representative systems, organized by agent roles and representative systems. *Agents as Simulations* model scientific processes, while *Agents as Tools* support empirical and analytical workflows.

System	SciSci Task	Key Mechanisms
Single-Agent Systems		
The AI Scientist (Lu et al., 2024)	End-to-end research	Iterative ideation, experimentation, and writing via single LM controller
ResearchAgent (Baek et al., 2025)	Idea generation	Iterative refinement over literature with entity-centric knowledge graph
SciMON (Wang et al., 2024b)	Novelty optimization	Retrieval-based inspiration; iterative comparison against prior work
CoQuest (Liu et al., 2024)	Research question generation	Breadth-first / depth-first RQ exploration with single LLM agent
REMOR (Taechoyotin and Acuna, 2025)	Peer review generation	Reasoning LLM with multi-objective RL; single model with GRPO training
Multi-Agent Systems		
VirSci (Su et al., 2025)	Research ideation	Role-separated agents in propose–critique–revise cycles
ResearchTown (Yu et al., 2025)	Reading, writing, review	Agent–data graph; role-governed interactions in modeled community
CiteAgent (Ji et al., 2025)	Citation dynamics	Agents embedded in citation network; micro–macro measurement layer
AgentReview (Jin et al., 2024)	Peer review simulation	Multi-phase interactions; controlled visibility between roles
MARG (D'Arcy et al., 2024)	Review generation	Specialized reviewer agents (experiments, clarity, impact); internal discussion
AgentRxiv (Schmidgall and Moor, 2025)	Collaborative research	Shared preprint server; inter-laboratory agent collaboration
SciAgents (Ghafarollahi and Buehler, 2024)	Scientific discovery	Multi-agent graph reasoning over ontology-based knowledge graphs
Agent Laboratory (Schmidgall et al., 2025)	Research assistance	Multiple LLM agents as coordinated research assistants
Hybrid Systems		
PaperQA2 (Skarlinski et al., 2024)	Literature QA	RAG pipeline with staged retrieval, ranking, and synthesis
Scideator (Radensky et al., 2025)	Structured ideation	Facet extraction + recombination; RAG modules for novelty checking
Agentic AutoSurvey (Liu et al., 2025b)	Survey synthesis	Role-specialized pipeline: search, cluster, draft, evaluate stages
Auto Research Vision (Liu et al., 2025a)	Full research workflow	Staged workflow (review $\rightarrow$ ideation $\rightarrow$ experiment $\rightarrow$ writing) with composable tools
AI Co-Scientist (Gottweis et al., 2025)	Hypothesis generation	Multi-stage pipeline with tournament-based ranking and validation
SciSciGPT (Shao et al., 2025)	SciSci analysis	Human-AI collaboration with staged tool-augmented workflows
ChemCrow (M. Bran et al., 2024)	Chemistry tasks	LLM + 18 chemistry tools; ReAct-style orchestration

Table 6: Classification of SciSci agent systems by architectural pattern. **Single-agent** systems use a single LM controller with iterative reasoning. **Multi-agent** systems distribute responsibilities across role-specialized agents with explicit coordination. **Hybrid** systems combine staged workflows, retrieval augmentation, and tool orchestration, often with elements of both single and multi-agent designs.