
Position: AI Agents in Scientific Teams Should Be Studied as Human-Agent Systems

Patrick Emami*, Jared D. Willard, Saumya Sinha, Truc Nguyen, Andrew Glaws
National Laboratory of the Rockies

Sameera Horawalavithana†, Gihan Panapitiya, Bruno Jacob, Siddhisanket Raskar
Pacific Northwest National Laboratory

Nithin Somasekharan, Ling Yue, Shaowu Pan
Rensselaer Polytechnic Institute

Brian Lu, Jason Eisner
Johns Hopkins University

Abstract

Large language model-based agents are increasingly deployed as collaborators in scientific discovery yet most current work focuses on the autonomous capabilities of “AI Scientists”. We argue that this overlooks the social aspects of scientific teamwork, and that studying AI Scientists as human-agent systems (HAS)—where the unit of analysis is the human-agent pair—is both underexplored and undervalued. We establish these points through literature and empirical analysis, and highlight recent incidences and studies which show that deploying agents in science without accounting for human-agent dynamics introduces near-term risks, including reduced diversity of scientific inquiry. Through analysis of real-world case studies, we show that scientists and agents can augment each other’s capabilities. We call for new research that adopts the HAS lens to develop mathematical frameworks for understanding and fostering human-AI synergy in scientific discovery.

1 Introduction

Modern science is a social endeavor that increasingly involves working in teams (Fortunato et al., 2018; NASEM, 2025). Collaboration helps scientists achieve their lofty goals by bringing experts with diverse scientific backgrounds and experiences together, combining their strengths to collectively answer highly complex questions. Science teams vary widely in their form and function, from small university laboratories to large institutions that coordinate hundreds of members. Meta-analysis of scientific teamwork has found that working in a team increases impact; in fact, teams are 6.3 times more likely to produce a highly cited paper than a solo author (Wuchty et al., 2007).

Today, scientific teamwork is undergoing a major shift as large language model (LLM)-based agents—systems that perceive, reason, and act in an environment, which we refer to simply as “agents”—are being deployed not only as tools for scientists, but as collaborators in scientific processes. The introduction of agents into scientific teamwork sits within broader trends of 1) increasing prevalence and advantage of scientific teamwork (Hall et al., 2018) and 2) increasing levels of automation in science. Position papers envision AI Scientists as evolving towards higher levels of autonomy with less and less reliance on human judgment and oversight (Gao et al., 2024; Liu et al., 2025; Zhang et al., 2025b;a; Wei et al., 2025a; Xie et al., 2025; Zhou et al., 2025a; Zheng et al., 2025; Ren et al., 2025; Tie et al., 2026). In Section 2, we analyze current agentic systems adapted for scientific discovery, which we call **AI Scientists**, finding that the majority of recent work only considers a narrow form of

*patrick.emami@nlr.gov

†yasanka.horawalavithana@pnnl.gov

Table 1: A representative set of agents categorized by the primary channel for human feedback (row-color). The three feedback dimensions (Type, Granularity, and Phase) follow Zou et al. (2025a)’s taxonomy of human feedback in LLM HAS. Systems can support multiple channels to varying degrees (see Table 2). See Section 2 for definitions.

System	Type	Granularity	Phase
Artifact-level			
AI Scientist v2 (Yamada et al., 2025)	Evaluative	Coarse	Post-task
Kosmos (Mitchener et al., 2025)	Evaluative	Coarse	Post-task
CycleResearcher (Weng et al., 2025a)	Evaluative	Coarse	Post-task
VirSci (Su et al., 2025)	Evaluative	Coarse	Post-task
AgentRxiv (Schmidgall & Moor, 2025)	Evaluative	Coarse	Post-task
AI-Researcher (Tang et al., 2026)	Evaluative	Coarse	Post-task
DeepScientist (Weng et al., 2025b)	Evaluative	Coarse	Post-task
Discovery-phase-level			
InternAgent (Team et al., 2025)	Evaluative	Coarse	Mid-task
Co-Scientist (Gottweis et al., 2025)	Evaluative	Coarse	Mid-task
AgentLaboratory (Schmidgall et al., 2025)	Evaluative	Coarse	Mid-task
Denario (Villaescusa-Navarro et al., 2025)	Corrective	Fine	Mid-task
Research-plan-level			
MAPPs (Zhou et al., 2025b)	Guidance	Fine	Mid-task
El Agente Q (Zou et al., 2025b)	Guidance	Fine	Pre-task
Organa (Darvish et al., 2025)	Guidance	Fine	Pre-task
SciSciGPT (Shao et al., 2025b)	Guidance	Fine	Pre-task
Asynchronous steering & interruption			
FreePhDLabor (Li et al., 2025)	Corrective	Fine	Mid-task
EvoScientist (Lyu et al., 2026)	Corrective	Fine	Mid-task
OmniScientist (Shao et al., 2025a)	Evaluative	Coarse	Mid-task
ScienceClaw (Wang et al., 2026)	Guidance	Coarse	Mid-task

human participation: setting goals and verifying work. Fewer systems support interaction geared at sustained, iterative, and joint collaboration typical of human-human teamwork. This literature has largely been isolated from the socio-technical realities of science.

If AI Scientists are to be effectively integrated into scientific teams, it is timely that we also begin to study them as human-agent systems (HAS). An HAS considers as core components: agents, human participants, an interface that facilitates human-agent interaction through actions, and the environment which the HAS is situated in (Haupt & Brynjolfsson, 2025; Zou et al., 2025a; Luo et al., 2026). In HAS, the unit of study is the pair: *humans and agents*. **Our position is that studying AI Scientists as HAS is underexplored and undervalued, and that new research which adopts the lens of HAS is needed to develop an understanding of human-AI synergy in scientific discovery settings.**

We argue that adopting an HAS lens can help us better understand—and thus potentially mitigate—near-term risks of deploying AI Scientists at scale. Evidence of the risks of AI in science is accruing; for example, a recent study showed that AI-augmented research produces three times as many papers at a 5% reduction in the diversity of explored topics (Hao et al., 2026). In Section 3, we provide further evidence from recent incidents and studies, and discuss the importance of framing AI Scientists as HAS to mitigation efforts.

We also argue that an HAS perspective presents a path to exploring new scientific questions and discovering new solutions inaccessible to humans alone and AI alone. In Section 4, we analyze recent real-world case studies of scientists co-working alongside agents that qualitatively illustrate how humans and agents have augmented each other’s outputs, achieving a level of human-AI synergy that tangibly improved scientific outcomes. Finally, we raise new questions about the study of agents in scientific discovery motivated by the HAS lens in Section 5, and discuss alternative views in Section 6.

Table 2: A preliminary evaluation of feedback-seeking behavior in a ReAct GPT-5-mini agent on the End-to-End Discovery (E2E) AstaBench benchmark. We compare three conditions: (1) the AstaBench planning-centric ReAct agent harness, (2) adding an ask_user tool and the HAS human feedback taxonomy to the system prompt as *feedback encouragement*, and (3) adding explicit guidance to always call the ask_user tool before every action. Condition (3) acts as an upper bound on agent disengagements. We use LLM-as-a-judge to classify questions under the taxonomy and show dimension-wise percentages (counts across 10 E2E development set tasks normalized by total questions asked), averaged across three judges for robustness. **Key findings:** The agent without explicit per-step instruction *does not know when to ask*, e.g., it prefers to make assumptions about how to handle ambiguous situations when critical resources are unavailable. With per-step enforcement, the dominant feedback type is evaluative, generally occurring mid-task. Second most common behavior is guidance-based, e.g., asking the user their preference given options, occurring heavily in the pre-task phase. Corrective feedback is rarely requested.

Condition			ask_user calls				Type				Phase			Granularity			E2E Score
ask_user tool	Feedback taxonomy	Per-step		E	G	C	I	D	P	C	F						
			0	0	0	0	0	0	0	0	0						0.39
✓	✓		1	0	1.0	0	1.0	0	0	0	1.0						0.36
✓	✓	✓	92	0.71	0.25	0.04	0.35	0.45	0.20	0.68	0.32						0.24

2 AI Scientists as HAS: A Review and Preliminary Experiment

We adopt an HAS lens to review the literature on AI Scientists, with the goal of broadly understanding how current systems interact with human scientists. Our review categorizes AI Scientist systems based on the primary channel which human feedback enters the scientific discovery loop: artifact-level feedback (Section 2.1), discovery phase-level feedback (Section 2.2), research planning feedback (Section 2.3), and asynchronous steering & interruption (Section 2.4). We adopt three human feedback dimensions from Zou et al. (2025a)’s taxonomy for LLM HAS in Table 1 to further organize systems: **Type**—*Evaluative* (E): assessment of output quality, *Guidance* (G): instructions, demonstrations, or critiques, or *Corrective* (C): user edits or fixes, **Granularity**—*Coarse* (C): Single assessment for an entire output or outcome, *Fine* (F): Targeted step-wise feedback, and **Phase**—*Initial setup & goal* (I): Pre-task, *During task execution* (D): Mid-task, *Post-task eval & refinement* (P).

2.1 Artifact-level feedback

Early AI Scientist systems automate the full discovery pipeline—ideation, experimentation, manuscript preparation—requiring humans only to specify a research question and evaluate final outputs (Lu et al., 2024; Yamada et al., 2025; Ghafarollahi & Buehler, 2025b; Tang et al., 2026; Weng et al., 2025b). Tree-search and Bayesian-optimization extensions improve exploration and reduce hallucinations by filtering hypotheses with computational validation (Yamada et al., 2025; Weng et al., 2025b; Liu et al., 2024b; Yu et al., 2025). Tighter internal feedback loops feed experiment results back into ideation (Yuan et al., 2025) or pair the agent with a peer-review trained reviewer model for iterative refinement (Weng et al., 2025a), reducing reliance on human review before final evaluation. Multi-agent variants organize LLM agents into collaborative teams (Su et al., 2025), structured multi-stage workflows (Tang et al., 2026), and swarm reasoning over knowledge graphs (Ghafarollahi & Buehler, 2025a); agent preprint servers enable sharing results for cross-lab iteration (Schmidgall & Moor, 2025). Autoresearch demonstrates that a simple coding harness and LLM can serve a similar role when also equipped with domain-specific tools and an iterative evaluation loop (Karpathy, 2025). In all cases, humans remain as initiator and final evaluator.

2.2 Discovery-phase-level feedback

A large class of AI Scientists permit humans to intervene after one or more phases of a sequential discovery workflow, canonically: literature review, ideation and hypothesis generation, experimentation, then manuscript preparation and review. Across these systems,

human feedback is primarily evaluative and corrective; granularity level varies from coarse (fixed review gates) to fine (editable phase-level memory).

Multi-agent systems assign phase-specific roles to agents and expose phase-level review gates where scientists provide high-level feedback (Schmidgall et al., 2025; Gottweis et al., 2025; Ghareeb et al., 2026; Villaescusa-Navarro et al., 2025). Denario’s Markdown-based memory system exposes each phase-specific module’s inputs and outputs, enabling fine-grained oversight by editing intermediate state before resuming execution (Villaescusa-Navarro et al., 2025). Transparency and rewind mechanisms (e.g., linking generated figures to source code, reverting to earlier phases) further increase steerability (Ifargan et al., 2025; Wang et al., 2026; Liu et al., 2026).

Phase-specific feedback targets common failure modes such as hypothesis vagueness. Strategies range from high-level textual feedback on ideas (Team et al., 2025; Gottweis et al., 2025; Ni et al., 2024), to fine-grained feedback on complex outputs like 3D crystal structures (Qi et al., 2024), to idea down-selection from ranked pools (Yamada et al., 2025; Jansen et al., 2025). Google’s Co-Scientist for hypothesis generation adds a generate-debate-evolve tournament over hypotheses, with scientists supplying objectives upfront and steering between rounds—yielding experimentally validated discoveries in drug repurposing and antimicrobial resistance (Gottweis et al., 2025).

2.3 Research-plan-level feedback

Many scientific activities do not fit the sequential workflows of artifact-level and phase-level systems, which offer limited visibility into low-level agent actions and restrict human involvement to fixed feedback gates. Planning-centric AI Scientists address this via a plan-critique-execute paradigm, generating high-level plans and incorporating human feedback before execution (He et al., 2026). Domain-specific systems constrain the planning action space to a narrow set of tools and workflows (Zou et al., 2025b; Huang et al., 2025; Darvish et al., 2025; Shao et al., 2025b; Zhou et al., 2025b), with feedback granularity ranging from pre-execution plan critique (Darvish et al., 2025) to per-step approval (Zhou et al., 2025b). The human role here is primarily *guidance oriented*: scientists proactively shape the agent’s course of action through collaborative planning rather than reacting to outputs.

2.4 Feedback via asynchronous steering & interruption

A minority of AI Scientists offer fine-grained, asynchronous human interaction during phased discovery workflows. FreePhDLabor (Li et al., 2025) supports real-time interruption by writing corrective feedback into persistent agent memory. ScienceClaw (Wang et al., 2026) enables scientists to inspect and redirect agents by commenting on the Infinite platform. EvoScientist (Lyu et al., 2026) adopts a human-on-the-loop paradigm with self-evolving memory that learns user preferences across sessions. OmniScientist (Shao et al., 2025a) aims at seamless human participation across pipeline phases through a collaborative research protocol, peer-review and the human-voted ScienceArena platform. Coding harnesses also support mid-run interruption, but only with continuous monitoring or agents that know when to ask for help (Table 2) an open challenge even for frontier LLMs (Gulati et al., 2026). Asynchronous steering and interruption offers the highest level of human control by enabling intervention at arbitrary points rather than only at phase or output boundaries.

2.5 Preliminary empirical support

Our synthesis of the literature reveals that most AI Scientists assign humans supervisory roles—scoping, auditing, approval—*treating discovery as an isolated optimization problem weakly tied to human-led research practices*. Artifact-level feedback is the most numerous; discovery-phase-level systems that gate input at workflow boundaries form the next tier. Planning-centric systems constitute a smaller but growing class, and only a handful support flexible asynchronous interaction.

To strengthen these findings, we also conduct a preliminary experiment on a representative planning-based AI Scientist, the ReAct-based agent harness and GPT-5-mini from

the AstaBench AI Scientist benchmark suite (Bragg et al., 2025)). We compare variants provided with an ask_user tool (Gulati et al., 2026) and increasing levels of feedback guidance. Questions are responded with an automatic continuation prompt (the AstaBench default). The results on 10 End-to-End Discovery tasks in Table 2 suggest there is room for improving the ability of agents to understand when to ask for feedback and how to ask for fine-grained, corrective help in challenging, end-to-end discovery settings. We expand later on the potential benefits of humans and AI Scientists mutually correcting each other’s mistakes in Section 4. While 10 tasks may appear nominally small, each task represents a full, computationally intensive idea-to-report discovery workflow, providing a deep rather than broad measure of agent behavior.

3 Studying AI Scientists as HAS can assist near-term risk mitigation

Deploying AI Scientists in scientific teams raises pressing epistemic, existential, and ethical questions. We argue that a HAS lens helps center human participation in discovery loops, aiding near-term risk mitigation. While future AI Scientists may transcend current limitations, our emphasis is on near-term risks exacerbated when humans offload deliberative processes to AI rather than participating as teammates.

3.1 Can AI Scientists reliably produce accurate and unbiased scientific knowledge?

LLMs notoriously **hallucinate** (Xie et al., 2025; Zhang et al., 2025a; Liu et al., 2025): a post-hoc review found 51 *accepted* NeurIPS 2025 papers containing 100 confirmed hallucinated citations¹; ACL 2026 similarly flagged over 100 accepted papers citing non-existent literature². More broadly, LLMs and agents are prone to fabricating data or results when faced with demanding tasks (Miyai et al., 2026; Luo et al., 2025) (also observed in our E2E experiment in Table 2). A related concern is **bias proliferation**: LLMs may over-represent certain paradigms while under-representing novel ideas (Liu et al., 2025). GPT-4-generated college admissions essays were collectively less novel than human-written ones (Moon et al., 2025), with worrisome implications for homogenization of scientific inquiry (Messeri & Crockett, 2024; Hao et al., 2026).

An HAS lens calls for systems that maintain human oversight in the near term to detect hallucinations and biases. End-of-process review (Section 2.1) is inadequate; asynchronous interaction at multiple granularities—including audits of trace logs and code—can surface issues earlier (Luo et al., 2025), if accounting for cognitive load introduced by the speed and scale of AI.

3.2 How will AI Scientists affect our scientific institutions and human expertise?

AI Scientists’ output volume strains **peer review**, which is already overloaded: NeurIPS and ICLR submissions increased $10.4\times$ from 2014 to 2024 without commensurate reviewer growth (Wei et al., 2025b; Xie et al., 2025; Zhang et al., 2025b). They also expand the **attack surface** of scientific software and data. Backdoor attacks can induce LLMs to produce malicious code, and poisoning crowd-sourced knowledge bases like Wikipedia can enable such attacks (Souly et al., 2025; Carlini et al., 2024). Perhaps most consequentially, over-reliance may cause **de-skilling** in scientists, especially among junior researchers (Xie et al., 2025; Zhang et al., 2025a). A recent study shows dependence on AI coding assistants reduces conceptual understanding (Shen & Tamkin, 2026). Younger users show higher susceptibility to cognitive offloading (Gerlich, 2025).

The HAS lens pushes us to maintain institutional health as we deploy AI Scientists. Automated governance—red teaming (Ge et al., 2024), adversarial evaluations, monitoring, and provenance tracking—can address misuse and security. The de-skilling question points more directly to HAS design: *how we interact with AI systems matters for skill formation* (Shen

¹<https://gptzero.me/news/neurips/>

²https://2026.aclweb.org/acl_statement/

& Tamkin, 2026), making HAS research an important lever for preserving the deliberate practice through which scientists develop competencies.

3.3 Who is responsible when AI Scientists cause harm or enable misuse?

When AI Scientists facilitate research in harmful or sensitive areas, **accountability** is unclear. NIST recognizes AI as uniquely exacerbating risks of enhancing chemical, biological, radiological, or nuclear weapons capabilities (AI, 2024). Legal liability for AI harms remains rapidly evolving. LLM-based systems also heighten concerns about **unintentional plagiarism**: even when text is not verbatim, determining originality is difficult. An expert review found 36% of LLM and AI Scientist-generated papers contained noticeable plagiarism (Gupta & Pruthi, 2025). Research on whether human involvement at consequential decision points preserves accountability (Rahwan et al., 2019) would benefit both scientific quality and safety. Designing human-agent interfaces with natural intervention opportunities and plagiarism detection tools that do not overburden peer review are important topics for future study.

4 The potential of human-AI synergy in science: case studies

We now turn to human-AI synergy, a general goal of any HAS. In HAS, information and capability asymmetries are a key driver of effective collaboration (Hemmer et al., 2025), creating opportunities for humans and AI to complement each other and outperform either member alone. Prior work has formalized synergy as complementary team performance in classification decision support, combining human and AI predictions to reduce errors (Bansal et al., 2021; Steyvers et al., 2022). Empirically, complementarity appears rarely (Vaccaro et al., 2024), though this may reflect structural barriers: incentives favor autonomy research over HAS, and costly human studies limit rigorous evaluation (Haupt & Brynjolfsson, 2025). A weaker but still important form is human augmentation, where AI assists people without requiring proof that AI cannot do better alone. Both complementarity and augmentation have the potential to greatly benefit scientific discovery.

Each case study explored next is drawn from a published account of a working scientist collaborating with an agent on a concrete scientific task. Together they illustrate patterns of mutual augmentation: humans contributing conceptual framing, domain judgment, and epistemic verification; agents compressing implementation while also contributing at the conceptual level.

4.1 Case study #1: “Vibe-coding” a minor complexity-theory result with Gemini 3 Pro

This case study, obtained from the collection by Woodruff et al. (2026), concerns a complexity theorist collaborating with Gemini 3 Pro inside Google Antigravity (a $\text{\LaTeX}$ -integrated IDE) to write up a theorem they had not previously had time to publish.

How AI augmented humans: Given an initial prompt to plan and draft the paper in the human author’s writing style, Gemini generated a structured roadmap (plan.md) and a draft manuscript (paper.tex), then iteratively expanded proofs and exposition over eight prompt turns. This reduced the overhead of assembling the proof details on paper plus drafting and organizing a publishable write-up of a result that might otherwise have remained informal.

How humans augmented AI: Human expertise was decisive at key steering points. Early in the process, the human expert identified missing technical context (including relevant prior work and a search-vs.-decision framing) and corrected an erroneous assumption in a lemma. The interaction succeeded because an expert could diagnose omissions and provide high-quality feedback; without that expertise, the same workflow would likely fail or produce a misleading paper.

The synergy: The workflow resembles an expert-plus-junior-collaborator pattern, where the human sets the research direction and quality standards while the agent accelerates drafting and execution. The case illustrates synergy that enabled a “minor-result” publication, but it also surfaces a risk raised in Section 3: if widely adopted, this interaction mode may

homogenize scientific writing and may be especially risky for early-career scientists who still need to develop core writing and reasoning skills through deliberate practice.

4.2 Case study #2: AI-assisted reduced physics modeling of thermonuclear burn propagation with GPT-5

This case study, from [Bubeck et al. \(2025\)](#), concerns an expert collaborating with GPT-5 through the ChatGPT interface to develop a reduced physics model of thermonuclear burn initiation and propagation. The exercise spanned four linked activities: (1) setting up a simplified PDE-based model for physical intuition, (2) implementing numerical simulation, (3) designing numerical experiments to expose key sensitivities, and (4) developing a theory to explain the simulation outcomes. The expert reported that the strongest outcome of this exercise was not time saving; rather it was (4) a theory-level explanation tying known physical relations into a closed-form predictor, used to verify that the numerically elicited trends were physically coherent.

How AI augmented humans: GPT-5 helped translate a human-provided concept into a concrete computational workflow. It drafted a PDE representation, incorporated domain-familiar physical ingredients, discretized the system, and scaffolded an optimization routine.

How humans augmented AI: Human expertise was essential in the numerics-heavy phase, where physically meaningful parameterization required iterative debugging. The human had to manually tune hot-spot and cold-fuel conditions, conductivities, alpha-particle stopping assumptions, and profile designs to obtain physically interesting behavior. Crucially, the human also provided epistemic checks: identifying when outputs were still noise, catching oversimplifications, and rejecting silent substitutions of full numerical solves with coarse approximations.

The synergy: This interaction resembled expert-plus-accelerated-expert-implementer collaboration. Work that might otherwise have required prolonged coordination across multiple human experts was compressed into a single focused session. However, GPT-5 exhibited familiar failure modes as Case #1: premature confidence and smoothing over thorny issues, requiring expert oversight to address.

5 Human-AI synergy in scientific discovery: open questions

Can complementary scientific team performance lead to discoveries that neither humans nor agents would have made working alone? This section builds on the risks identified in Section 3 and case studies (Section 4) to motivate new research questions.

To help guide this discussion, we introduce a toy model of the overall utility obtained by a HAS on a scientific discovery task defined within a simple decision theoretic framework for human-agent collaboration inspired by [Shao et al. \(2024\)](#). The utility of the team with humans H and agents A consists of just two terms: the **collaboration advantage** $CA(H, A)$ and the **collaboration disadvantage** $CD(H, A)$. $CA(H, A)$ is an advantage function that measures the team performance against the best member of the team, (i.e., *complementarity*, similar to the *synergy gap* in multi-agent teams ([Pappu et al., 2026](#))). The term $CD(H, A) \geq 0$ measures the cost of collaborating, capturing additional costs paid by the team due to communication and coordination overhead (e.g., tokens spent on messages, additional cognitive load incurred by human team members). Our toy model of overall utility

$$U(H, A) = CA(H, A) - CD(H, A) \quad (1)$$

suggests a goal of maximizing $CA(H, A)$ while minimizing $CD(H, A)$.

5.1 Which factors of scientific discovery environments increase CA or reduce CD?

This question explores how attributes of scientific discovery environments relate to the overall utility $U(H, A)$ achieved by a fixed scientific team. Looking at the environments in our case studies (Section 4), we start to see common factors.

- Task outputs tend to be subjective research artifacts such as papers, theorems, or novel abstractions. Here, a human expert’s ability to quickly judge the work, built up over years of experience, is valuable.
- Tasks tend to have long-horizons and are open-ended and curiosity-driven, consisting of sub-tasks that produce research artifacts which may initially appear weakly related to the original desired outcome.
- Tasks may rely on limited partial observations to infer underlying structure.

These desiderata contrast with scientific environments where CD is likely to overwhelm CA: tasks where agent trajectories and outputs can be automatically verified (Ye et al., 2026) or tasks where abundant training data exists (e.g., protein structure prediction (Abramson et al., 2024)). Understanding which factors impact overall utility at the environment level has implications for benchmark design as well as downstream applications.

5.2 Which dimensions of human-agent scientific teamwork increase CA or reduce CD?

LLMs have changed how scientists interface with AI dramatically: humans and LLMs fluidly communicate in natural language, and agents can display human-like reasoning and planning. They can also be configured to emulate distinct cognitive and social behaviors, particularly through role playing, elevating their status from tools to collaborators. We offer the following questions that explore dimensions of scientific teamwork which impact utility:

- Trust is essential in human-AI interaction yet is dynamic and multifaceted. How should we design interactions to establish and maintain trust between humans and agent teammates in scientific contexts, and how does trust impact utility $U(H, A)$?
- What is the relationship between AI over-reliance and utility $U(H, A)$ (Section 3.2), and how should we design for interactions that minimize over-reliance risks?
- Does integrating agents into real-time collaborative processes via asynchronous communication (e.g., AI processing delays (Liu et al., 2024a)) and mixed-initiative interaction (Horvitz, 1999) increase CA at a reasonable increase in CD?

6 Alternative Views

The primary opposing view to our core position is that a HAS perspective may provide little practical value for scientific teams once full costs are tallied. Even if utility improves for some settings, the evaluation complexity and cost (not captured by collaboration disadvantage CD) may offset benefits. *Centaur evaluations* (Haupt & Brynjolfsson, 2025) that compare HAS against humans and agents alone are important yet currently expensive and difficult to reproduce. The idea is that the HAS is justified only when $U(H, A) - \max(U(H, H), U(A, A))$ is clearly large enough to exceed measurement costs.

We acknowledge the real challenge of cost-effective measurement of HAS in scientific discovery, particularly given the open-ended and long-horizon nature of these tasks. However, a reasonable path is staged evaluation: start with lower cost evaluations (surveys and questionnaires, focused unit-tests of individual scientific collaborative capabilities (Somasekharan et al., 2026), observational data like transcripts of agent logs and aggregate statistics, and case studies), then invest in centaur studies when signals are strong. We also propose the adoption of lightweight proxy metrics aligned with our utility model. To measure CD, researchers can evaluate **Review Time** as a cognitive load proxy, where a judge LLM measures how concise and actionable are the agent’s questions and how long and meandering are research artifacts, as well as **Intervention Density**, which measures edit distance between agent artifacts and a human’s finalized version. To measure CA, we advocate for **Simulated Oracle Synergy**, inspired by prior work on simulated feedback in multi-turn settings (Wang et al., 2024). Here a stronger LLM acts as a proxy human scientist, allowing researchers to automatically measure how effectively an agent ingests and applies critical feedback.

7 Conclusion

As we deploy AI agents into scientific teams, we should study these agents not just as solving an isolated optimization problem but also as a member of a *human-agent system*. This lens centers research on sustaining human participation throughout discovery workflows, enabling fine- and coarse-grained intervention beyond artifact review and fixed gates, and offers concrete pathways to mitigate near-term risks while cultivating human-AI synergy in science. Our evidence motivates a mathematical framework which encourages exploring interaction modes that maximize the collaboration advantage or lower coordination costs. Studying AI Scientists as HAS is a pragmatic route to achieving safe and effective human-agent co-discovery.

Acknowledgments

This work was authored in part by the National Laboratory of the Rockies for the U.S. Department of Energy (DOE), operated under Contract No. DE-AC36-08GO28308. It was also supported in part by the Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05-76RLO1830. This material is based upon work supported by the U.S. Department of Energy, Office of Science, ASCR under Award Number DE-SC0025425. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DOE, the United States Government or any agency thereof. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. This paper has been cleared by PNNL for public release as PNNL-SA-219899.

References

- Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016): 493–500, 2024.
- NIST AI. Artificial intelligence risk management framework: Generative artificial intelligence profile. *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA*, 2024.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021. doi: <https://doi.org/10.1145/3411764.3445717>.
- Jonathan Bragg, Mike D’Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefan Candra, Jason Dunkelberger, Dan Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Cecile Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. Astabench: Rigorous benchmarking of ai agents with a scientific research suite, 2025. URL <https://arxiv.org/abs/2510.21652>.
- Sébastien Bubeck, Christian Coester, Ronen Eldan, Timothy Gowers, Yin Tat Lee, Alexandru Lupascu, Mehtaab Sawhney, Robert Scherrer, Mark Sellke, Brian K Spears, et al. Early science acceleration experiments with gpt-5. *arXiv preprint arXiv:2511.16072*, 2025.

-
- Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning Web-Scale Training Datasets is Practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425, May 2024. doi: 10.1109/SP54263.2024.00179. URL <https://ieeexplore.ieee.org/abstract/document/10646610>. ISSN: 2375-1207.
- Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, et al. Organa: A robotic assistant for automated chemistry experimentation and characterization. *Matter*, 8 (2), 2025.
- Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.
- Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.
- Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabisa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. MART: Improving LLM safety with multi-round automatic red-teaming. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1927–1937, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.107. URL <https://aclanthology.org/2024.naacl-long.107/>.
- Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025.
- Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22): 2413523, 2025a.
- Alireza Ghafarollahi and Markus J Buehler. Sparks: Multi-agent artificial intelligence model discovers protein design principles. *arXiv preprint arXiv:2504.19017*, 2025b.
- Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, pp. 1–3, 2026.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.
- Anmol Gulati, Hariom Gupta, Elias Lumer, Sahil Sen, and Vamse Kumar Subbiah. Ask early, ask late, ask right: When does clarification timing matter for long-horizon agents? *arXiv preprint arXiv:2605.07937*, 2026.
- Tarun Gupta and Danish Pruthi. All that glitters is not novel: Plagiarism in ai generated research. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25721–25738, 2025.
- Kara L Hall, Amanda L Vogel, Grace C Huang, Katrina J Serrano, Elise L Rice, Sophia P Tsakraklides, and Stephen M Fiore. The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American psychologist*, 73(4):532, 2018.
- Qianyu Hao, Fengli Xu, Yong Li, and James Evans. Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Nature*, pp. 1–7, 2026.

-
- Andreas Haupt and Erik Brynjolfsson. Position: Ai should not be an imitation game: Centaur evaluations. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.
- Zeyu He, Hannah Kim, Dan Zhang, and Estevam Hruschka. How to steer your multi-agent system: Human-llm collaborative planning. *arXiv preprint arXiv:2605.23023*, 2026.
- Patrick Hemmer, Max Schemmer, Niklas Köhl, Michael Vössing, and Gerhard Satzger. Complementarity in human-ai collaboration: Concept, sources, and evidence. *European Journal of Information Systems*, 34(6):979–1002, 2025.
- Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999.
- Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.
- Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous llm-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1): AIoa2400555, 2025.
- Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13370–13467, 2025.
- Andrej Karpathy. autoresearch. <https://github.com/karpathy/autoresearch>, 2025. 2025a.
- Ed Li, Junyu Ren, Xintian Pan, Cat Yan, Chuanhao Li, Dirk Bergemann, and Zhuoran Yang. Build your personalized research group: A multiagent framework for continual and interactive science automation. *arXiv preprint arXiv:2510.15624*, 2025.
- Fan Liu, Jindong Han, Tengfei Lyu, Weijia Zhang, Zhe-Rui Yang, Lu Dai, Cancheng Liu, and Hao Liu. Foundation models for scientific discovery: From paradigm enhancement to paradigm transition. *Authorea Preprints*, 2025.
- Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, Xiaoyan Bai, Haizhong Zheng, et al. The last human-written paper: Agent-native research artifacts. *arXiv preprint arXiv:2604.24658*, 2026.
- Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–25, 2024a.
- Zijun Liu, Kaiming Liu, Yiqi Zhu, Xuanyu Lei, Zonghan Yang, Zhenhe Zhang, Peng Li, and Yang Liu. AIGS: Generating Science from AI-Powered Automated Falsification, November 2024b. URL <http://arxiv.org/abs/2411.11910>. arXiv:2411.11910 [cs].
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.
- Hanjun Luo, Chiming Ni, Jiaheng Wen, Zhimu Huang, Yiran Wang, Bingduo Liao, Sylvia Chung, Yingbin Jin, Xinfeng Li, Wenyuan Xu, XiaoFeng Wang, and Hanan Salam. Centaureval: Benchmarking human-in-the-loop value in agentic coding. In *Forty-third International Conference on Machine Learning*, 2026.
- Ziming Luo, Atoosa Kasirzadeh, and Nihar B Shah. The more you automate, the less you see: Hidden pitfalls of ai scientist systems. *arXiv preprint arXiv:2509.08713*, 2025.

-
- Yougang Lyu, Xi Zhang, Xinhao Yi, Yuyue Zhao, Shuyu Guo, Wenxiang Hu, Jan Piotrowski, Jakub Kaliski, Jacopo Urbani, Zaiqiao Meng, et al. Evoscientist: Towards multi-agent evolving ai scientists for end-to-end scientific discovery. *arXiv preprint arXiv:2603.08127*, 2026.
- Lisa Messeri and MJ Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.
- Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, et al. Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.
- Atsuyuki Miyai, Mashiro Toyooka, Takashi Otonari, Zaiying Zhao, and Kiyoharu Aizawa. Jr. AI scientist and its risk report: Autonomous scientific exploration from a baseline paper. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL <https://openreview.net/forum?id=0eV062d8Sw>.
- Kibum Moon, Adam E Green, and Kostadin Kushlev. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, pp. 100207, 2025.
- NASEM. *The Science and Practice of Team Science*. The National Academies Press, Washington, DC, 2025. ISBN 978-0-309-73441-7. doi: 10.17226/29043. URL <https://nap.nationalacademies.org/catalog/29043/the-science-and-practice-of-team-science>.
- Ziqi Ni, Yahao Li, Kaijia Hu, Kunyuan Han, Ming Xu, Xingyu Chen, Fengqi Liu, Yicong Ye, and Shuxin Bai. Matpilot: an llm-enabled ai materials scientist under the framework of human-machine collaboration. *arXiv preprint arXiv:2411.08063*, 2024.
- Aneesh Pappu, Batu El, Hancheng Cao, Carmelo di Nolfo, Yanchao Sun, Meng Cao, and James Zou. Multi-agent teams hold experts back. *arXiv preprint arXiv:2602.01011*, 2026.
- Jingyuan Qi, Zian Jia, Minqian Liu, Wangzhi Zhan, Junkai Zhang, Xiaofei Wen, Jingru Gan, Jianpeng Chen, Qin Liu, Mingyu Derek Ma, Bangzheng Li, Haohui Wang, Adithya Kulkarni, Muhao Chen, Dawei Zhou, Ling Li, Wei Wang, and Lifu Huang. Metascientist: A human-ai synergistic framework for automated mechanical metamaterial design, 2024. URL <https://arxiv.org/abs/2412.16270>.
- Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.
- Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025.
- Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.320. URL <https://aclanthology.org/2025.findings-emnlp.320/>.
- Chenyang Shao, Dehao Huang, Yu Li, Keyu Zhao, Weiquan Lin, Yining Zhang, Qingbin Zeng, Zhiyu Chen, Tianxing Li, Yifei Huang, et al. Omniscientist: Toward a co-evolving ecosystem of human and ai scientists. *arXiv preprint arXiv:2511.16931*, 2025a.

-
- Erzhuo Shao, Yifang Wang, Yifan Qian, Zhenyu Pan, Han Liu, and Dashun Wang. Scis-cigpt: Advancing human-ai collaboration in the science of science. *arXiv preprint arXiv:2504.05559*, 2025b.
- Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*, 2024.
- Judy Hanwen Shen and Alex Tamkin. How ai impacts skill formation. *arXiv preprint arXiv:2601.20245*, 2026.
- Nithin Somasekharan, Youssef Hassan, Shiyao Lin, Gihan Panapitiya, Patrick Emami, Anurag Acharya, Sameera Horawalavithana, and Shaowu Pan. Sciconvbench: Benchmarking llms on multi-turn clarification for task formulation in computational science. *arXiv preprint arXiv:2605.18630*, 2026.
- Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougán, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples, October 2025. URL <http://arxiv.org/abs/2510.07192>. arXiv:2510.07192 [cs].
- Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022.
- Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, et al. 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)*, pp. 28201–28240, 2025.
- Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520, 2026.
- InternAgent Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Runmin Ma, Yusong Hu, Zhiyin Yu, Xiaohan He, Songtao Huang, et al. Internagent: When agent becomes the scientist–building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*, 2025.
- Guiyao Tie, Jiawen Shi, Dingjie Song, Yixiao Huang, Ziji Sheng, Xueyang Zhou, Daizong Liu, Pan Zhou, Yongchao Chen, Ran Xu, et al. Autoresearch ai: Towards ai-powered research automation for scientific discovery. *arXiv preprint arXiv:2605.23204*, 2026.
- Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, 2024.
- Francisco Villaescusa-Navarro, Boris Bolliet, Pablo Villanueva-Domingo, Adrian E Bayer, Aidan Acquah, Chetana Amancharla, Almog Barzilay-Siegal, Pablo Bermejo, Camille Bilodeau, Pablo Cárdenas Ramírez, et al. The denario project: Deep knowledge ai agents for scientific discovery. *arXiv preprint arXiv:2510.26887*, 2025.
- Fiona Y Wang, Lee Marom, Subhadeep Pal, Rachel K Luu, Wei Lu, Jaime A Berkovich, and Markus J Buehler. Autonomous agents coordinating distributed discovery through emergent artifact exchange. *arXiv preprint arXiv:2603.14312*, 2026.
- Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. In *International Conference on Learning Representations*, volume 2024, pp. 32593–32627, 2024.

-
- Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntao Cao, et al. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111*, 2025a.
- Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. The ai imperative: Scaling high-quality peer review in machine learning. *arXiv preprint arXiv:2506.08134*, 2025b.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cyclereviewer: Improving automated research via automated review. In *International Conference on Learning Representations*, volume 2025, pp. 3669–3709, 2025a.
- Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. Deepscientist: Advancing frontier-pushing scientific findings progressively. *arXiv preprint arXiv:2509.26603*, 2025b.
- David P Woodruff, Vincent Cohen-Addad, Lalit Jain, Jieming Mao, Song Zuo, Mohammad Hossein Bateni, Simina Branzei, Michael P Brenner, Lin Chen, Ying Feng, et al. Accelerating scientific research with gemini: Case studies and common techniques. *arXiv preprint arXiv:2602.03837*, 2026.
- Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.
- Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, et al. How far are ai scientists from changing the world? *arXiv preprint arXiv:2507.23276*, 2025.
- Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.
- Haotian Ye, Haowei Lin, Jingyi Tang, Yizhen Luo, Caiyin Yang, Chang Su, Rahul Thapa, Rui Yang, Ruihua Liu, Zeyu Li, et al. Evaluation-driven scaling for scientific discovery. *arXiv preprint arXiv:2604.19341*, 2026.
- Zhaojian Yu, Kaiyue Feng, Yilun Zhao, Shilin He, Xiao-Ping Zhang, and Arman Cohan. Alpharesearch: Accelerating new algorithm discovery with language models. *arXiv preprint arXiv:2511.08522*, 2025.
- Jiakang Yuan, Xiangchao Yan, Bo Zhang, Tao Chen, Botian Shi, Wanli Ouyang, Yu Qiao, Lei Bai, and Bowen Zhou. Dolphin: moving towards closed-loop auto-research through thinking, practice, and feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21768–21789, 2025.
- Haoxuan Zhang, Ruochi Li, Yang Zhang, Ting Xiao, Jiangping Chen, Junhua Ding, and Haihua Chen. The evolving role of large language models in scientific innovation: Evaluator, collaborator, and scientist. *arXiv preprint arXiv:2507.11810*, 2025a.
- Pengsong Zhang, Heng Zhang, Huazhe Xu, Renjun Xu, Zhenting Wang, Cong Wang, Animesh Garg, Zhibin Li, Arash Ajoudani, and Xinyu Liu. Scaling laws in scientific discovery with ai and robot scientists. *arXiv preprint arXiv:2503.22444*, 2025b.
- Tianshi Zheng, Zheyang Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259*, 2025.
- Lianhao Zhou, Hongyi Ling, Cong Fu, Yepeng Huang, Michael Sun, Wendi Yu, Xiaoxuan Wang, Xiner Li, Xingyu Su, Junkai Zhang, et al. Autonomous agents for scientific discovery: Orchestrating scientists, language, code, and physics. *arXiv preprint arXiv:2510.09901*, 2025a.

-
- Lianhao Zhou, Hongyi Ling, Keqiang Yan, Kaiji Zhao, Xiaoning Qian, Raymundo Arróyave, Xiaofeng Qian, and Shuiwang Ji. Toward greater autonomy in materials discovery agents: Unifying planning, physics, and scientists. *arXiv preprint arXiv:2506.05616*, 2025b.
- Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, et al. Llm-based human-agent collaboration and interaction systems: A survey. *arXiv preprint arXiv:2505.00753*, 2025a.
- Yunheng Zou, Austin H Cheng, Abdulrahman Aldossary, Jiaru Bai, Shi Xuan Leong, Jorge Arturo Campos-Gonzalez-Angulo, Changhyeok Choi, Cher Tian Ser, Gary Tom, Andrew Wang, et al. El agente: An autonomous agent for quantum chemistry. *Matter*, 8 (7), 2025b.