Position: AI Scientists Are Not Yet Ready for Open-Ended and Fully Autonomous Scientific Discovery

Anonymous ACL submission

Abstract

We argue that **current AI scientist systems are not yet ready for open-ended and fully autonomous scientific discovery**. Despite impressive capabilities in automating research workflows, these systems produce research-like artifacts rather than validated science—optimizing for surface plausibility while lacking the judgment, creativity, and real-world grounding essential to genuine discovery. Through systematic analysis and human evaluation, we identify three critical gaps: (1) the *real-world environment gap*—absence of infrastructure for validating AI-generated hypotheses against physical reality; (2) the *professional skills gap*—lack of deep domain expertise beyond general-purpose reasoning; and (3) the *quality verification gap*—lack of scalable mechanisms for ensuring that AI-generated scientific claims are reliable, reproducible, and scientifically verifiable. We propose corresponding directions: scaling verifiable real-world research environments, cultivating domain-specific agent skills, and developing reliability-aware frameworks. Until these fundamental gaps are bridged, AI scientists should serve as collaborative partners amplifying human capabilities, not as autonomous researchers.

1 Introduction

Scientific discovery has powered human civilization for centuries, yet the enterprise faces unprecedented strain. Scientific publication output grows at roughly 4% per year (Bornmann et al., 2021), while peer review systems buckle under increasing pressure (Hanson et al., 2024). Major machine learning conferences now receive over **20,000 submissions annually**, and the reproducibility crisis undermines trust in published findings (Gundersen et al., 2018). Against this backdrop, AI scientist, which we refer to as *agentic systems driven by Large Language Models (LLMs)*, have emerged

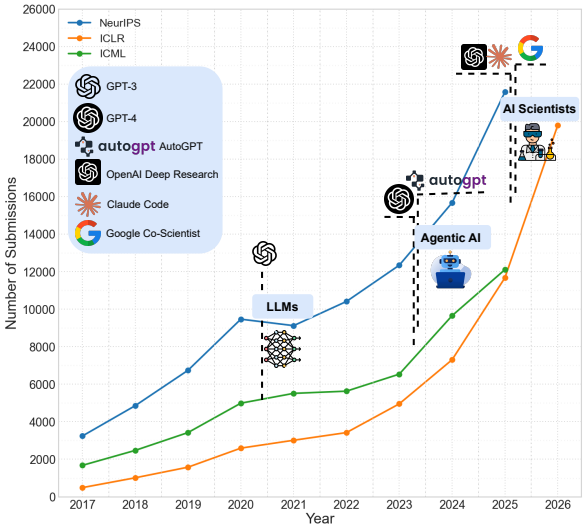

Figure 1: **AI-Driven Research Explosion Crisis.** The rapid evolution of AI-powered research tools has been accompanied by a surge in conference submissions, while the quality and reproducibility of AI-assisted research outputs remain highly inconsistent. (Data from `https://papercopilot.com/statistics`)

with the promise of automating research at scale: generating hypotheses, designing experiments, and producing manuscripts at very low cost (Lu et al., 2026; Yamada et al., 2025; Gottweis et al., 2025).

The adoption of AI scientists is rapid and widespread (Lu et al., 2024; Xie et al., 2025b). As Figure 1 illustrates, AI-powered research tools, ranging from LLMs to agentic systems and AI scientists, have proliferated dramatically (Jiang et al., 2025a), contributing to the rapid growth of AI-generated scientific content and increasing pressure on existing quality-control mechanisms. Our human study of 25 active AI researchers (Appendix E) confirms this trend: **80% report using AI systems frequently** for research tasks, with coding assistance (100%), paper writing (80%), and literature review (76%) as dominant use cases. In response, major scientific platforms such as arXiv have re-

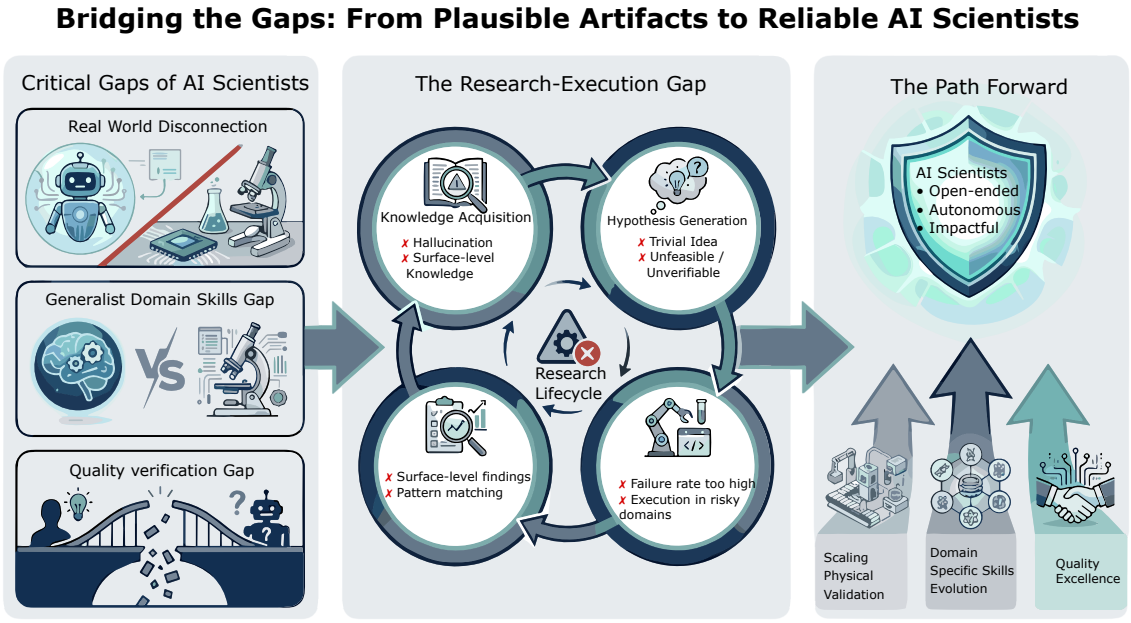

Figure 2: **Overview of our paper**. We analyze the current state of AI scientists, identify three critical gaps (real-world environment, professional skills, quality verification), examine deficiencies across the research pipeline, and propose directions for developing AI scientists as collaborative partners rather than autonomous replacements.

cently introduced penalties targeting low-quality or unverified AI-generated submissions, including hallucinated references and fabricated experimental claims.[1] Understanding the capabilities and limitations of AI scientists is therefore not merely academic—it is essential for the research community navigating this technological transformation.

Yet despite impressive technical achievements, a sobering reality emerges: current AI scientist systems primarily produce research-like artifacts, such as codebases, trained models, and full research papers, rather than independently validated scientific knowledge (Kusumegi et al., 2025; Beel et al., 2025). They optimize for surface plausibility—well-structured papers, fluent prose, comprehensive citations—while struggling with the deeper requirements of scientific validity: hypotheses grounded in causal mechanisms, experiments that genuinely test claims, and conclusions that withstand independent scrutiny (Zhu et al., 2025c). The gap between *appearing* scientific and *being* scientifically valid represents the central challenge.

> Our claim: Current AI scientists are not yet ready for open-ended, fully autonomous scientific discovery.

[1] See: https://www.nature.com/articles/d41586-026-01595-5

Bridging three critical gaps is essential before these systems can fulfill their transformative potential (Figure 2). The path forward requires not simply scaling existing approaches, but systematically addressing fundamental limitations in how AI scientists interact with the real world, acquire domain expertise, and collaborate with human researchers:

Real-world environment gap. Current AI scientists operate primarily in sandboxed computational environments, lacking infrastructure to validate discoveries against physical reality. Genuine scientific impact requires scalable, verifiable real-world research environments where AI-generated hypotheses can be tested and confirmed through authentic experimentation.

Professional skills gap. General-purpose language models lack the specialized domain expertise that human scientists accumulate over careers. Moving beyond shallow pattern matching to deep scientific reasoning demands cultivating professional agent skills—domain-specific knowledge, methodological rigor, and judgment calibrated to particular research contexts.

Scientific quality gap. Current AI scientist systems increasingly optimize for artifact production, without corresponding guarantees of scientific validity, reproducibility, or verification. As

autonomous research generation scales, the field risks accelerating the production of plausible but insufficiently validated scientific artifacts. The field should prioritize reliable and verifiable scientific workflows over unconstrained autonomous research generation.

Future AI scientists must operate within scalable real-world research environments that enable physical validation, develop modular professional skills grounded in domain expertise and scientific workflows, and incorporate reliability-aware discovery mechanisms that prioritize verification, traceability, and oversight. In this paper, we argue that advancing AI scientists will require tightly integrating environment scaling, professional scientific skills, and reliable discovery frameworks into the next generation of scientific agent systems.

> Until these fundamental gaps are bridged, AI scientists should serve as collaborative partners amplifying human capabilities, not as autonomous replacements.

2 The Current State: Capabilities and Limitations

The term open-endedness has been widely studied in artificial intelligence and reinforcement learning (Hughes et al., 2024), where it typically refers to a system's ability to continuously generate novel and learnable outputs through interaction with an environment.

In this paper, we define **open-ended scientific discovery** as the ability of an AI scientist system to autonomously pursue scientific research in unbounded and evolving environments, where research goals and methodologies are not predefined . Unlike narrow scientific tasks with fixed objectives or workflows, open-ended discovery requires continuously adapting to new problems, unexpected observations, and changing research directions.

We define a **fully autonomous AI scientist** as a system capable of independently conducting the full scientific discovery cycle, including problem formulation, hypothesis generation, experimentation, validation, and iterative refinement, without human involvement in core scientific decisions.

2.1 What Current AI Scientists Can Do

The past two years have witnessed remarkable progress in automated scientific research systems. These advances span the entire research pipeline, demonstrating capabilities that challenge assumptions about the boundaries of machine intelligence in science.

Literature synthesis AI systems now efficiently process vast literature corpora, extracting key findings and identifying potential research gaps. Tools leveraging retrieval-augmented generation synthesize knowledge across thousands of papers, surfacing connections that might escape individual researchers (He et al., 2025; Jin et al., 2024; Wei et al., 2024). Specialized literature agents demonstrate proficiency in organizing scientific knowledge and tracking evolving consensus (Li et al., 2024; Wang et al., 2024b).

Hypothesis generation. Large language models exhibit surprising facility for generating research hypotheses. Studies show that LLM-based systems can produce plausible, sometimes validated, scientific hypotheses from background information alone (Qi et al., 2023; Zhou et al., 2024). Multi-agent frameworks where proposing and reviewing agents iterate on ideas yield increasingly refined proposals (Baek et al., 2024; Wang et al., 2024a; Su et al., 2024). Recent work even suggests AI-generated ideas can match human novelty in controlled settings (Si et al., 2024).

Experimental design and execution. In computational domains, AI agents automate experimental workflows with increasing sophistication. Code generation systems translate high-level descriptions into executable experiments (Chen et al., 2021; Austin et al., 2021; Li et al., 2022; Miao et al., 2025), while orchestration frameworks coordinate multiple specialized tools (Shen et al., 2023; M. Bran et al., 2024). Self-driving laboratories extend these capabilities to physical experimentation in chemistry and materials science (Boiko et al., 2023; Steiner et al., 2019; Dai et al., 2024; Darvish et al., 2025).

2.2 Where Current Systems Fall Short

Despite these capabilities, systematic evaluation reveals fundamental limitations that prevent current AI scientists from achieving genuine scientific impact. To ground our analysis empirically, we conducted a human study surveying 25 researchers from the Computer Science and AI community—all with hands-on AI tool experience and academic publication backgrounds (see Appendix E for full methodology and results). These empirical findings contextualize three critical gaps we identify below.

Real-world environment gap. A persistent disconnect exists between computational predictions and real-world outcomes (Lombardo et al., 2021). Current AI scientists operate primarily in sandboxed computational environments, lacking infrastructure to validate discoveries against physical reality. In materials science, molecules that appear promising in computational screening often fail in synthesis (Tshitoyan et al., 2019). The AI Scientist system reports experiment success rates below 50%, with code that executes but implements algorithms different from those described (Lu et al., 2024). Reproducibility benchmarks reveal that even state-of-the-art systems struggle to faithfully execute described procedures (Zhao et al., 2025; Siegel et al., 2024; Han et al., 2025). While pioneering systems like Coscientist (Boiko et al., 2023) and mobile robotic platforms (Dai et al., 2024; Darvish et al., 2025) demonstrate physical-world integration, these remain isolated examples requiring substantial human infrastructure. Scaling verifiable real-world research environments remains an open challenge.

Professional skills gap. General-purpose language models, despite their breadth, lack the specialized domain expertise that human scientists accumulate over careers. Analysis reveals that AI-generated literature reviews demonstrate "breadth without depth"—comprehensive citation without genuine intellectual engagement (Martin-Boyle et al., 2024). Generalization studies show systematic biases in how AI systems summarize and synthesize scientific findings (Peters and Chin-Yee, 2025; Ektefaie et al., 2024). A particularly concerning manifestation is citation hallucination: recent analysis reveals that fabricated references have infiltrated published literature at alarming rates (Sakai et al., 2026). These hallucinations appear credible to casual inspection but corrupt the scientific record. Moving beyond shallow pattern matching to deep scientific reasoning demands cultivating domain-specific agent skills—methodological rigor and judgment calibrated to particular research contexts.

Quality verification gap. Current AI scientist systems can generate research artifacts at unprecedented scale, yet ensuring scientific quality remains fundamentally unresolved. Recent systems such as AI Scientist-v2 (Yamada et al., 2025), Kosmos (Mitchener et al., 2025), and AI-Researcher (Tang et al., 2026) primarily optimize hypothesis generation, code execution, and manuscript production, while reproducibility studies continue to reveal implementation errors, hallucinated citations, weak experimental validation, and unreliable conclusions (Siegel et al., 2024; Starace et al., 2025; Sakai et al., 2026). This imbalance between generation and verification risks accelerating low-quality research exploration beyond the capacity of existing scientific quality-control pipelines. AI scientists therefore require scalable infrastructures for reproducibility, validation, traceability, and scientific accountability, rather than stronger generation capabilities alone.

3 More Gaps Across the Research Pipeline

Following the capability-level framework proposed by recent surveys (Xie et al., 2025a), we analyze the concrete gaps across five core stages of scientific research: (1) knowledge acquisition, (2) hypothesis generation, (3) experimental design and execution, (4) analysis and interpretation, and (5) validation and evolution. At each stage, current AI scientists exhibit systematic deficiencies that prevent reliable scientific discovery.

3.1 Knowledge Acquisition

Knowledge acquisition requires retrieving and comprehending domain-specific scientific literature (Xie et al., 2025a). Current systems demonstrate impressive breadth in literature processing, but remain limited in depth of understanding.

Precision-recall trade-offs in retrieval. Current AI scientist systems heavily rely on keyword-based matching, often retrieving classic but potentially outdated papers while missing recent advances (He et al., 2025; Jiang et al., 2025b). The enormous scale and diversity of scientific literature makes it difficult to accurately retrieve the most relevant and up-to-date information for a given research problem. Both precision and recall remain inadequate for high-stakes scientific applications.

Citation hallucination. A particularly concerning manifestation is the systematic fabrication of citations. Recent studies have identified hundreds of hallucinated references in major conference proceedings (Sakai et al., 2026). These hallucinations appear credible to casual inspection but corrupt the scientific record when subsequent work builds upon false foundations.

Table 1: Critical gaps across the scientific research pipeline.

Research stage	Core capability	Current gap
Knowledge acquisition	Literature retrieval and understanding	Inaccurate retrieval and hallucinated citations undermine reliable scientific grounding (He et al., 2025; Jiang et al., 2025b; Sakai et al., 2026).
Hypothesis generation	Novel and feasible scientific idea generation	Generated hypotheses are often repetitive, weakly grounded, or experimentally infeasible (Qi et al., 2023; Zhou et al., 2024; Lu et al., 2024).
Experimental design and execution	Experiment implementation and execution	Current systems frequently fail to produce reliable, reproducible, and intention-aligned experiments (Chan et al., 2024; Starace et al., 2025; Siegel et al., 2024).
Analysis and interpretation	Scientific reasoning and interpretation	AI systems struggle with causal reasoning, methodological judgment, and scientifically meaningful interpretation (Ektefaie et al., 2024; Peters and Chin-Yee, 2025).
Validation and evolution	Verification, quality control, and iterative improvement	Validation infrastructure cannot keep pace with AI-generated research outputs, creating major quality and reliability risks (Gundersen et al., 2018; Hanson et al., 2024; Weng et al., 2025).

3.2 Hypothesis Generation

Hypothesis generation represents the key feature distinguishing AI scientist systems from automated scientific tools (Xie et al., 2025a). Yet current systems face fundamental challenges in producing hypotheses that are simultaneously novel, feasible, and scientifically grounded.

The novelty-feasibility trade-off. Studies show that while LLMs can generate novel hypotheses (Qi et al., 2023; Zhou et al., 2024; Si et al., 2024), these ideas are often less feasible or disconnected from experimental reality. The core challenge is that high-quality hypothesis generation requires domain intuition—understanding which questions are tractable, scientifically meaningful, and worth pursuing.

Repetition and lack of originality. Empirical evidence suggests that ideas produced by AI systems tend to lack true originality and are frequently repetitive across different runs or even across different models (Lu et al., 2024; Xie et al., 2025a). LLMs are fundamentally limited by their training data, constraining their ability to move beyond well-trodden conceptual ground.

3.3 Experimental Design and Execution

The capability to design, implement, and execute experiments transforms an AI scientist from an idea generator into an autonomous scientific intelligence (Xie et al., 2025a). However, this stage reveals the most severe capability gaps.

Implementation and execution failures. Benchmark evaluations show that even state-of-the-art LLMs struggle to translate conceptual understanding into reliable experimental implementations (Chan et al., 2024; Starace et al., 2025; Siegel et al., 2024). The AI Scientists reports experiment success rates below 50%, with implementations often deviating from the intended algorithms (Lu et al., 2024). These failures include incorrect hyperparameters, flawed preprocessing pipelines, and inconsistent statistical analyses (Miao et al., 2025). As AI-assisted "vibe coding" becomes increasingly common, such errors may propagate into published computational research (Siegel et al., 2024; Lu et al., 2024).

Physical world constraints. In physical sciences, the gap widens further. Self-driving laboratories demonstrate impressive automation in chemistry and materials science (Boiko et al., 2023; Dai et al., 2024; Darvish et al., 2025), but remain confined to narrow and pre-structured problem spaces (Stach et al., 2021). Molecules that appear promising in computational screening often fail in synthesis (Tshitoyan et al., 2019), while real-world execution introduces complications absent from simulation environments (Matsiko, 2024). This disconnect highlights the persistent real-world environment gap.

3.4 Analysis and Interpretation

Data analysis in scientific domains requires not only computational proficiency but scientific judgment in distinguishing signal from noise, identifying patterns, and drawing valid inferences.

Correlation without causation. AI systems excel at identifying statistical patterns but struggle with scientific interpretation. They may detect correlations without distinguishing causation, apply sophisticated methods without checking assumptions, or produce analytically convincing outputs that overlook methodological flaws (Ektefaie et al., 2024; Peters and Chin-Yee, 2025).

Visualization and communication gaps. While AI systems can generate plots and tables, they often fail to select informative visualizations (Liang and You, 2025) or emphasize scientifically meaningful results.

3.5 Validation and Evolution

The most severe gap concerns the asymmetry between generation and validation. AI scientists gen-

erate claims far faster than they can be verified, creating systemic risks for scientific integrity.

The validation bottleneck. No scalable infrastructure exists for systematic verification of AI-generated hypotheses. Human expert review cannot keep pace with AI generation, while existing evaluations show that AI-generated papers often suffer from experimental weaknesses, methodological ambiguity, and limited novelty (Weng et al., 2025; Xie et al., 2025a). This reflects the unresolved quality verification gap.

Lack of evolutionary capability. A mature AI scientist should continuously advance its research abilities based on feedback (Xie et al., 2025a). Current systems predominantly focus on single-task completion rather than managing long-term research cycles with comprehensive planning and iteration. When relying solely on self-generated feedback through internal reflection, AI scientists are prone to "looping errors"—mistakes amplified over multiple iterations rather than corrected.

AI review quality crisis. AI-generated reviews often provide generic feedback and fail to identify substantive methodological issues (Weng et al., 2025). Evidence from Agents4Science further shows persistent gaps in scientific judgment when AI agents act as both authors and reviewers (Bianchi et al., 2025).

4 Safety: Capability Without Safeguards

As AI scientists become more capable, safety concerns intensify across multiple dimensions (Tang et al., 2025).

Technical vulnerabilities. Current systems exhibit critical vulnerabilities including susceptibility to prompt injection attacks, memory contamination that introduces fabricated citations, and unsafe tool operations in laboratory settings (Zhu et al., 2025b). Benchmarks like AgentHarm reveal that even aligned models comply with harmful requests in multi-step scientific workflows (Andriushchenko et al., 2024).

Coordination risks. Multi-agent AI scientist systems introduce additional risks: small communication errors can cascade into large-scale reasoning failures (Ghafarollahi and Buehler, 2024; Song et al., 2025). The lack of standardized communication protocols for scientist-to-scientist interactions results in inefficient information exchange and suboptimal integration of external criticism.

Systemic concerns. Safety frameworks remain underdeveloped relative to capability advances, creating risks of dual-use research, biased outputs that distort scientific priorities, and over-standardization that suppresses creative inquiry (Stahl, 2021; Kowald et al., 2024). Without robust ethical constraints, AI scientist systems may autonomously enter dangerous research domains, accelerating the development of potentially harmful technologies before adequate safeguards can be implemented (Xie et al., 2025a).

5 Proposed Directions

5.1 Scaling Real-World Research Environments

AI-generated hypotheses require physical validation. Recent advances in autonomous laboratories (Steiner et al., 2019; Dai et al., 2024; Darvish et al., 2025; Angelopoulos et al., 2024) and robotic experimentation (Matsiko, 2024; Zhang et al., 2025b) have begun enabling closed-loop scientific workflows in which AI-generated hypotheses can be automatically tested and iteratively refined. Emerging frameworks further emphasize environment scaling as a key path toward general agentic intelligence by evaluating agents in increasingly diverse and dynamic settings (Zhang et al., 2025a; Liu et al., 2025; Wen et al., 2025; Fang et al., 2025; Froger et al., 2025). However, existing systems remain fragmented and lack scalable infrastructure for reliable real-world scientific discovery.

Future progress may require treating scientific discovery as an embodied interaction with physical environments rather than a purely computational reasoning task. One important direction is the development of standardized interfaces connecting AI planning systems with laboratory equipment, robotic platforms, simulation engines, and scientific instruments (Stach et al., 2021). AI scientists may additionally require persistent closed-loop research environments where hypotheses, experimental feedback, and environmental observations continuously inform subsequent reasoning and experimentation. More broadly, scalable benchmark ecosystems grounded in real-world outcomes—such as experimental reproducibility, successful synthesis, and adaptive interaction with evolving environments—may become essential for training and evaluating scientific agents beyond static computational tasks (Siegel et al., 2024).

5.2 Cultivating Professional Agent Skills for Science

Recent work has begun augmenting general-purpose LLMs with domain-specific capabilities through specialized scientific training (Gao et al., 2024; Zheng et al., 2025), tool integration with external software systems (M. Bran et al., 2024; Shen et al., 2023), and modular agent skill frameworks (Gottweis et al., 2025; Agent Skills, 2026). Domain-specific agents for materials science (Ni et al., 2024), chemistry (M. Bran et al., 2024), and biomedicine (Gao et al., 2024) further demonstrate the potential of incorporating procedural knowledge and specialized workflows into scientific agents. Meanwhile, self-evolving agent frameworks explore how agents can iteratively improve tools, memory, and workflows through interaction and feedback (Lin et al., 2024; Team et al., 2025; ang Gao et al., 2026; Ou et al., 2025; Zhai et al., 2025). Despite these advances, current systems remain far from expert-level scientific reasoning and continue to suffer from shallow understanding, cascading failures, and hallucinated outputs (Zhu et al., 2025a; Shao et al., 2025).

Future AI scientists may require professional skills that emerge from long-term interaction with domain-specific scientific processes rather than large-scale text prediction alone. One important direction is the development of training environments grounded in iterative experimentation, scientific feedback, failure analysis, and long-horizon research processes. AI scientists may additionally require persistent memory and experience accumulation mechanisms that allow agents to refine research strategies, track prior failures, and develop domain-specific judgment over extended time horizons. More broadly, modular scientific skill ecosystems capable of independently developing, evaluating, and composing specialized reasoning abilities—such as experimental design, statistical analysis, and scientific verification—may become essential for robust scientific discovery.

5.3 Reliable Scientific Discovery Guidance

Verification-first scientific pipelines. Future AI scientist systems should produce outputs that are reproducible, executable, traceable, and auditable by construction. Scientific claims should be accompanied by transparent reasoning traces, executable workflows, reproducible environments, and provenance records that enable independent verification (Schmidgall and Moor, 2025).

Scalable scientific validation infrastructure. Reliable AI-driven discovery requires scalable mechanisms for validating scientific outputs beyond sandboxed simulations. Emerging directions include autonomous laboratory platforms for closed-loop experimentation, reproducibility benchmarks across diverse environments, and distributed verification networks for cross-validating AI-generated hypotheses.

Reliability-aware oversight and scientific generation. Future AI scientist systems should incorporate uncertainty calibration, confidence estimation, self-verification, and execution-grounded evaluation into the generation process (Xie et al., 2025a). At the system level, robust oversight mechanisms—including AI disclosure policies, validation checkpoints, independent review processes, and provenance tracking—will be essential for maintaining scientific accountability and preventing unreliable claims from propagating through the literature (Feng et al., 2024).

6 Alternative Views

We acknowledge credible perspectives that challenge our position.

AI scientists are already achieving reliable discovery level performance. Systems like AlphaFold (Jumper et al., 2021) and Coscientist (Boiko et al., 2023) demonstrate genuine scientific contributions. Google's AI co-scientist has generated novel hypotheses subsequently validated experimentally (Gottweis et al., 2025). In materials science, AI-driven discovery platforms have identified novel compounds (Tshitoyan et al., 2019; Stach et al., 2021). The AI Scientist produces complete research papers autonomously (Lu et al., 2024), and NovelSeek demonstrates closed-loop hypothesis-to-verification workflows (Team et al., 2025). A related critique suggests our framework exaggerates the distance to reliable discovery—after all, language models have improved dramatically through scaling (Bang et al., 2023), and historical precedent suggests initial skepticism about AI capabilities often proves unfounded (Xu et al., 2021).

Our response: We acknowledge these achievements but note a crucial distinction: successful systems operate in well-defined problem spaces with clear validation criteria (Wang et al., 2023). AlphaFold predicts structures against experimentally determined ground truth; Coscientist executes pre-

defined reaction types in controlled settings. The gaps we identify—open-ended hypothesis generation, cross-domain reasoning, and autonomous validation—distinguish narrow successes from general scientific competence (Reddy and Shojaee, 2025). Our human study shows that while AI systems receive "Good" ratings for execution tasks (coding: 3.92/5, writing: 3.68/5), they are rated "Poor" for creative tasks (hypothesis generation: 2.71/5, reproduction: 2.63/5). The top limitations cited—hallucination (28%) and lack of creativity (24%)—persist across state-of-the-art systems (Lu et al., 2024; Siegel et al., 2024), suggesting fundamental limitations rather than incremental shortfalls. Bridging these gaps requires new capabilities: physical experimentation infrastructure that does not yet exist at scale (Stach et al., 2021), domain-specific skill acquisition beyond current training paradigms (Gao et al., 2024), and principled collaboration protocols that current architectures do not support (Gottweis et al., 2025). The reproducibility crisis in human science (Gundersen et al., 2018) demonstrates that even validated methodologies fail without proper infrastructure—scaling alone cannot substitute for structural investment.

7 Call to Action

Bridging the gaps requires coordinated investment across the research community.

For Researchers. Addressing the professional skills and quality verification gaps requires more transparent and systematic evaluation of AI scientific workflows. First, *releasing AI interaction logs*, including prompts, reasoning traces, and verification procedures, could help researchers analyze failure modes and improve specialized scientific agents. Second, the community would benefit from *shared repositories of failure cases*, including hallucinations and incorrect scientific conclusions. Finally, *paired evaluations against human baselines* may provide more reliable assessments of scientific progress, particularly when negative results and failure modes are openly reported.

For Institutions. Closing the real-world environment gap requires infrastructure for scalable validation and oversight of AI-assisted science. First, universities and research institutions could establish *shared autonomous experimentation platforms* connecting AI-generated hypotheses with robotic laboratories under controlled conditions. Second, institutions will need *stronger accountability mechanisms*, including standardized disclosure frameworks for AI involvement in scientific workflows. Finally, *oversight structures analogous to Institutional Review Boards* may become necessary to ensure sufficient human verification before AI-assisted research is published.

For Funding Agencies. Resolving the gaps requires sustained investment in scientific infrastructure and verification ecosystems for AI-assisted discovery. First, funding agencies should support *shared validation infrastructure*, including autonomous laboratories, verification networks, and benchmark ecosystems for AI-generated discoveries. Second, AI-science funding programs should require *explicit human verification protocols* with clear oversight checkpoints. Finally, *long-term collaboration between AI researchers and domain scientists* remains essential for developing robust domain-specific agent skills.

For the AI Community. Existing evaluations of AI scientists remain heavily benchmark-driven and often prioritize plausibility over scientific validity. First, the community should develop *domain-grounded benchmarks with real-world validation signals*, such as testing whether AI-proposed molecules can actually be synthesized rather than merely appearing plausible. Second, *modular scientific agent skills* that can be independently evaluated and transferred across domains may improve reliability and adaptability. Finally, *interpretable confidence estimation and uncertainty calibration* remain essential in scientific settings where errors may propagate into downstream research.

8 Conclusion

We argue that **current AI scientist systems are not yet ready for open-ended, fully autonomous scientific discovery** due to three critical gaps: the absence of real-world validation environments, the lack of specialized professional domain skills, and insufficient quality verification frameworks. These capability limitations, compounded by emergent problems including citation hallucination, low-quality AI reviews, and unverified code generation, create systemic risks for scientific integrity. We advocate for a strategic reorientation from pursuing fully autonomous AI scientists toward developing AI scientific partners that amplify human capabilities while preserving accountability—the path to AI-enabled discovery runs through deeper integration with human expertise, not increasing autonomy.

Limitations

This paper is primarily a position and synthesis work rather than a comprehensive empirical benchmark study. Our conclusions reflect an interpretation of current evidence on AI scientist systems and may evolve as the field progresses rapidly. While we ground our discussion in recent literature and a human evaluation study, our empirical analysis remains limited in scale and primarily reflects perspectives from researchers in computer science and AI-related fields. Broader interdisciplinary evaluations across domains such as biology, chemistry, physics, and medicine may reveal different capability profiles and deployment considerations.

In addition, our discussion focuses specifically on open-ended and fully autonomous scientific discovery rather than narrow or human-supervised scientific workflows. We do not deny that current AI systems already provide substantial value in assisted research settings, including coding, literature review, experimental automation, and hypothesis exploration. Some capability gaps identified in this paper may narrow substantially with future advances in reasoning, tool use, and autonomous experimentation. However, we argue that challenges related to scalable validation, scientific accountability, and reliable real-world experimentation are unlikely to be resolved through model scaling alone.

References

Agent Skills. 2026. Agent skills: Overview. https://agentskills.io/home. Accessed: 2026-01-28.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. 2024. Agentharm: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024*.

Huan ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, et al. 2026. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *Preprint*, arXiv:2507.21046.

Angelos Angelopoulos, James F Cahoon, and Ron Alterovitz. 2024. Transforming science labs into automated factories of discovery. *Science Robotics*, 9(95):eadm6991.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

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

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Joeran Beel, Min-Yen Kan, and Moritz Baumgart. 2025. Evaluating sakana's ai scientist: Bold claims, mixed results, and a promising future? In *ACM SIGIR Forum*, volume 59, pages 1–20. ACM New York, NY, USA.

Federico Bianchi, Owen Queen, Nitya Thakkar, Eric Sun, and James Zou. 2025. Exploring the use of ai authors and reviewers at agents4science. *Nature Biotechnology*, pages 1–4.

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

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. 2024. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374.

Tianwei Dai, Sriram Vijayakrishnan, Filip T Szczypiński, Jean-François Ayme, Ehsan Simaei, Thomas Fellowes, Rob Clowes, Lyubomir Kotopanov, Caitlin E Shields, Zhengxue Zhou, et al. 2024. Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, pages 1–8.

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

Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian G Marin, Marinka Zitnik, and Maha Farhat. 2024. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, 6(12):1512–1524.

Runnan Fang, Shihao Cai, Baixuan Li, Jialong Wu, Guangyu Li, Wenbiao Yin, Xinyu Wang, Xiaobin Wang, Liangcai Su, Zhen Zhang, et al. 2025. Towards general agentic intelligence via environment scaling. *arXiv preprint arXiv:2509.13311*.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large language model-based human-agent collaboration for complex task solving. *arXiv preprint arXiv:2402.12914*.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, et al. 2025. Are: Scaling up agent environments and evaluations. *arXiv preprint arXiv:2509.17158*.

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

Alireza Ghafarollahi and Markus J Buehler. 2024. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*.

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

Odd Erik Gundersen, Yolanda Gil, and David W Aha. 2018. On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications. *AI magazine*, 39(3):56–68.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. 2025. The personality illusion: Revealing dissociation between self-reports & behavior in llms. *arXiv preprint arXiv:2509.03730*.

Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, et al. 2025. Pasa: An llm agent for comprehensive academic paper search. *arXiv preprint arXiv:2501.10120*.

Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. 2024. Open-endedness is essential for artificial superhuman intelligence. *arXiv preprint arXiv:2406.04268*.

Pengcheng Jiang, Jiacheng Lin, Zhiyi Shi, Zifeng Wang, Luxi He, Yichen Wu, Ming Zhong, Peiyang Song, Qizheng Zhang, Heng Wang, et al. 2025a. Adaptation of agentic ai. *arXiv preprint arXiv:2512.16301*.

Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025b. s3: You don't need that much data to train a search agent via rl. *arXiv preprint arXiv:2505.14146*.

Qiao Jin, Robert Leaman, and Zhiyong Lu. 2024. Pubmed and beyond: biomedical literature search in the age of artificial intelligence. *EBioMedicine*, 100.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Dominik Kowald, Sebastian Scher, Viktoria Pammer-Schindler, Peter Müllner, Kerstin Waxnegger, Lea Demelius, Angela Fessl, Maximilian Toller, Inti Gabriel Mendoza Estrada, Ilija Šimić, et al. 2024. Establishing and evaluating trustworthy ai: overview and research challenges. *Frontiers in Big Data*, 7:1467222.

Keigo Kusumegi, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. 2025. Scientific production in the era of large language models. *Science*, 390(6779):1240–1243.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2024. Scilitllm: How to adapt llms for scientific literature understanding. *arXiv preprint arXiv:2408.15545*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Chumeng Liang and Jiaxuan You. 2025. Evaluating LLM-generated diagrams as graphs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12678–12690, Suzhou, China. Association for Computational Linguistics.

Guanyu Lin, Tao Feng, Pengrui Han, Ge Liu, and Jiaxuan You. 2024. Arxiv copilot: A self-evolving and efficient LLM system for personalized academic assistance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 122–130, Miami, Florida, USA. Association for Computational Linguistics.

Zichen Liu, Anya Sims, Keyu Duan, Changyu Chen, Simon Yu, Xiangxin Zhou, Haotian Xu, Shaopan Xiong, Bo Liu, Chenmien Tan, Chuen Yang Beh, Weixun Wang, Hao Zhu, Weiyan Shi, Diyi Yang, Michael Shieh, Yee Whye Teh, Wee Sun Lee, and Min Lin. 2025. Gem: A gym for agentic llms. *Preprint*, arXiv:2510.01051.

Teo Lombardo, Marc Duquesnoy, Hassna El-Bouysidy, Fabian Årén, Alfonso Gallo-Bueno, Peter Bjørn Jørgensen, Arghya Bhowmik, Arnaud Demortiere, Elixabete Ayerbe, Francisco Alcaide, et al. 2021. Artificial intelligence applied to battery research: hype or reality? *Chemical reviews*, 122(12):10899–10969.

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*.

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

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.

Anna Martin-Boyle, Aahan Tyagi, Marti A Hearst, and Dongyeop Kang. 2024. Shallow synthesis of knowledge in gpt-generated texts: A case study in automatic related work composition. *arXiv preprint arXiv:2402.12255*.

Amos Matsiko. 2024. Advancing scientific discovery with the aid of robotics.

Jiacheng Miao, Joe R. Davis, Yaohui Zhang, Jonathan K. Pritchard, and James Zou. 2025. Paper2agent: Reimagining research papers as interactive and reliable ai agents. *Preprint*, arXiv:2509.06917.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, et al. 2025. Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*.

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

Yixin Ou, Wangchunshu Zhou, Shengwei Ding, Long Li, Jialong Wu, Tiannan Wang, Jiamin Chen, Shuai Wang, Xiaohua Xu, Ningyu Zhang, et al. 2025. Symbolic learning enables self-evolving agents. *AI Open*.

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776.

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*.

Chandan K Reddy and Parshin Shojaee. 2025. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28601–28609.

David B Resnik and Susan A Elmore. 2016. Ensuring the quality, fairness, and integrity of journal peer review: A possible role of editors. *Science and Engineering Ethics*, 22:169–188.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. Hallucitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in acl conferences. *Preprint*, arXiv:2601.18724.

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

Shuai Shao, Qihan Ren, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, Dongrui Liu, et al. 2025. Your agent may misevolve: Emergent risks in self-evolving llm agents. *arXiv preprint arXiv:2509.26354*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

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*.

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

Peiyang Song, Pengrui Han, and Noah Goodman. 2025. A survey on large language model reasoning failures. In *2nd AI for Math Workshop@ ICML 2025*.

Eric Stach, Brian DeCost, A Gilad Kusne, Jason Hattrick-Simpers, Keith A Brown, Kristofer G Reyes, Joshua Schrier, Simon Billinge, Tonio Buonassisi, Ian Foster, et al. 2021. Autonomous experimentation systems for materials development: A community perspective. *Matter*, 4(9):2702–2726.

Bernd Carsten Stahl. 2021. Ethical issues of ai. In *Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies*, pages 35–53. Springer.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Sebastian Steiner, Jakob Wolf, Stefan Glatzel, Anna Andreou, Jarosław M Granda, Graham Keenan, Trevor Hinkley, Gerardo Aragon-Camarasa, Philip J Kitson, Davide Angelone, et al. 2019. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*, 363(6423):eaav2211.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *arXiv preprint arXiv:2410.09403*.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2026. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520.

Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. 2025. Risks of ai scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16(1):8317.

NovelSeek Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, et al. 2025. Novelseek: When agent becomes the scientist–building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

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

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

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. 2024b. Autosurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems*, 37:115119–115145.

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. 2024. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546.

Yule Wen, Yixin Ye, Yanzhe Zhang, Diyi Yang, and Hao Zhu. 2025. Real-time reasoning agents in evolving environments. *Preprint*, arXiv:2511.04898.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*.

Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, Jian Wu, and Yue Zhang. 2025a. How far are ai scientists from changing the world? *arXiv preprint arXiv:2507.23276*.

Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, et al. 2025b. How far are ai scientists from changing the world? *arXiv preprint arXiv:2507.23276*.

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

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

Yunpeng Zhai, Shuchang Tao, Cheng Chen, Anni Zou, Ziqian Chen, Qingxu Fu, Shinji Mai, Li Yu, Jiaji Deng, Zouying Cao, et al. 2025. Agentevolver: Towards efficient self-evolving agent system. *arXiv preprint arXiv:2511.10395*.

Jiayi Zhang, Yiran Peng, Fanqi Kong, Cheng Yang, Yifan Wu, Zhaoyang Yu, Jinyu Xiang, Jianhao Ruan, Jinlin Wang, Maojia Song, et al. 2025a. Autoenv: Automated environments for measuring cross-environment agent learning. *arXiv preprint arXiv:2511.19304*.

Pengsong Zhang, Heng Zhang, Huazhe Xu, Renjun Xu, Zhenting Wang, Cong Wang, Animesh Garg, Zhibin Li, Arash Ajoudani, and Xinyu Liu. 2025b. Scaling laws in scientific discovery with ai and robot scientists. *arXiv preprint arXiv:2503.22444*.

Xuanle Zhao, Zilin Sang, Yuxuan Li, Qi Shi, Shuo Wang, Duzhen Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Autoreproduce: Automatic ai experiment reproduction with paper lineage. *arXiv preprint arXiv:2505.20662*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pages 1–11.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. 2025a. Where llm agents fail and how they can learn from failures. *Preprint*, arXiv:2509.25370.

Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, Peixuan Han, Yue Su, Haofei Yu, and Jiaxuan You. 2025b. Safescientist: Toward risk-aware scientific discoveries by llm agents. *arXiv preprint arXiv:2505.23559*.

Minjun Zhu, Qiujie Xie, Yixuan Weng, Jian Wu, Zhen Lin, Linyi Yang, and Yue Zhang. 2025c. Ai scientists fail without strong implementation capability. *arXiv preprint arXiv:2506.01372*.

A Acknowledgements

We sincerely thank the researchers who participated in our survey. Their insights, experiences, and thoughtful feedback provided valuable perspectives that helped shape our understanding of current AI scientist systems and their limitations. We deeply appreciate their time and contributions to this work.

B Statement on LLM Usage

Large language models were used as assistive tools during the preparation of this paper. Their usage included supporting literature exploration, suggesting potentially relevant references, drafting and polishing prose, and brainstorming figure organization. All referenced works, factual claims, interpretations, and final manuscript decisions were independently verified and revised by the authors. The authors retained responsibility for the content, arguments, and conclusions presented in this paper.

C Open Questions

When are AI scientists appropriate to use? An important open question is identifying which stages of scientific research are suitable for AI scientists and which still fundamentally require human expertise. Current systems are particularly effective for accelerating repetitive and large-scale tasks, including literature retrieval, code generation, data processing, experiment automation, and broad hypothesis exploration (Lu et al., 2024). In these settings, AI systems can substantially improve efficiency by searching larger solution spaces and reducing manual workload. However, AI scientists remain unreliable for tasks requiring deep domain intuition, causal reasoning, long-horizon planning, and high-stakes scientific judgment. Our human study suggests researchers remain especially skeptical of AI-driven hypothesis generation, experiment design, and autonomous validation, where hallucinations and superficial reasoning can easily propagate into downstream research. We therefore argue that AI scientists are currently better viewed as acceleration tools for bounded and verifiable subtasks rather than autonomous researchers capable of independent scientific discovery. Determining clear boundaries for safe and effective deployment remains an important challenge for the field.

How far are we from reliable AI scientists? Our human study (Appendix E) shows clear capability gaps: AI scientists are rated below "Fair" ($<$3.0/5) on core creative tasks—hypothesis generation (2.71), reproduction (2.63), and experiment design (2.81). Main concerns are hallucination (28%) and lack of creativity (24%). 64% say early-stage human oversight is still essential; only 12% expect fully autonomous AI researchers. This indicates current AI systems are far from independent, reliable scientific discovery, due to lacking judgment, domain intuition, and creative insight (Si et al., 2024; Xie et al., 2025a).

Evaluation as a fundamental challenge. A key open problem is how to evaluate AI scientific contributions without known ground truth (Siegel et al., 2024). Existing benchmarks focus on isolated skills, but science demands integrated performance across the workflow. We argue evaluation should be broken down by process—assessing knowledge quality, hypothesis novelty, implementation, analysis, and validation—to pinpoint capability gaps and track true progress (Weng et al., 2025; Xie et al., 2025a).

Credit assignment and authorship. As AI systems increasingly contribute to research, fundamental questions arise about scholarly credit. Current authorship conventions assume human responsibility for claims (Resnik and Elmore, 2016), but how should contributions from AI systems be acknowledged? Should AI-generated hypotheses, code, or text receive formal attribution?

D Impact Statement

This paper aims to guide the responsible development and deployment of AI scientist systems. By identifying critical gaps between current capabilities and reliable scientific discovery, we hope to steer research investment toward addressing foundational limitations rather than prematurely scaling unreliable systems. The potential positive impact includes reducing the risk of scientific misinformation from AI-generated content, preserving human accountability in research, and ensuring AI systems genuinely advance rather than undermine scientific integrity. We acknowledge that overly cautious interpretations of our position could slow beneficial AI applications in science; however, we believe the greater risk lies in deploying systems that produce plausible but unverified claims at scale. Our advocacy for human-AI partnership frameworks reflects the conviction that the path to transformative AI-enabled discovery requires collaboration, not replacement.

E Human Study: AI Scientists Survey

E.1 Survey Overview and Methodology

We conducted an anonymous survey to understand how researchers perceive, use, and envision the future of AI Scientists. The survey was distributed within the Computer Science and AI research community in January, 2026.

Participant Recruitment and Ethics: All responses were collected anonymously from researchers with hands-on experience using AI-based research tools and academic research/publication experience. 100% of participants (N=25) explicitly agreed to allow their responses to be used for academic research purposes. No personally identifiable information was retained.

Sample Characteristics (N=25):

- **Research Domain:** 96% Computer Science, 4% Physics
- **Research Experience:** 60% with 1–3 years, 16% with 3–5 years, 12% with <1 year, 8% with 5–10 years, 4% with >10 years
- **AI Tool Usage:** 80% use AI systems frequently, 16% occasionally, 4% tried a few times

E.2 Survey Questions

The survey consisted of 15 questions across four sections: (1) *Demographics* (research domain, experience years); (2) *Current Usage* (usage frequency, scenarios, capability ratings, AI system types); (3) *Human-AI Collaboration* (involvement modes, critical stages, issue severity); and (4) *Future Perspectives* (limitations, collaboration vision, concerns, superhuman timeline). Questions included single-choice, multiple-choice, 5-point Likert scales, and open-ended responses. Full questionnaire available upon request.

E.3 Detailed Survey Results

E.3.1 AI System Usage Patterns

Usage Frequency: 80% of participants use AI systems *frequently* for research, indicating deep integration into research workflows. Figure 3 shows that coding/debugging is universally adopted (100%), followed by paper writing (80%) and literature review (76%). Hypothesis generation (32%) and paper reproduction (20%) see lower adoption.

Types of AI Systems: General-purpose LLMs (ChatGPT, Claude, Gemini) have universal adoption (100%), followed by coding agents (64%).

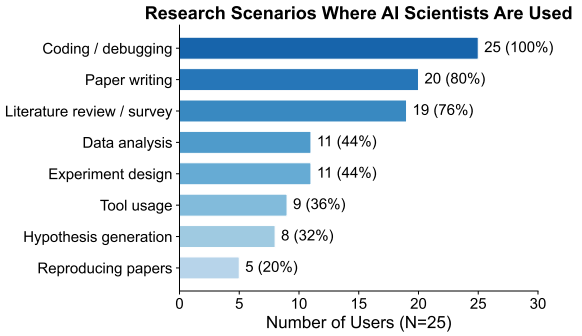

Figure 3: Research scenarios where AI scientists are used (N=25).

Fully autonomous research agents are used by only 20%.

E.3.2 Capability Assessment by Research Module

Figure 4 presents capability ratings across eight research modules:

- **High-rated (≥3.5/5):** Coding/debugging (3.92), Paper writing (3.68), Literature review (3.56), Data analysis (3.55)
- **Low-rated (<3.0/5):** Reproducing papers (2.63), Hypothesis generation (2.71), Experiment design (2.81)

This reveals a clear **execution-innovation gap**: AI excels at execution tasks but struggles with creative reasoning requiring deeper scientific understanding.

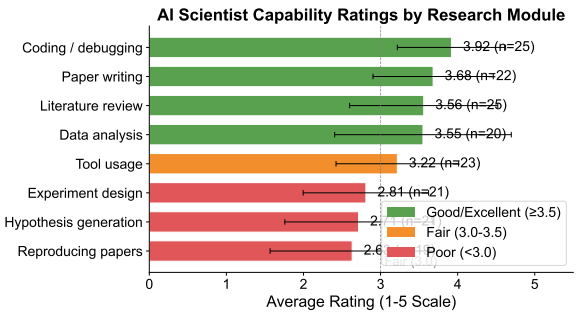

Figure 4: AI scientist capability ratings by research module (1-5 scale). Error bars show standard deviation.

E.3.3 Human-AI Collaboration Patterns

Figure 5 shows that 64% of researchers consider early stages (planning: 36%, task formulation: 28%) most critical for human involvement. Only 16% are comfortable with minimal human intervention.

Future Vision: The community is split: 44% envision equal partnership, 44% prefer human-led

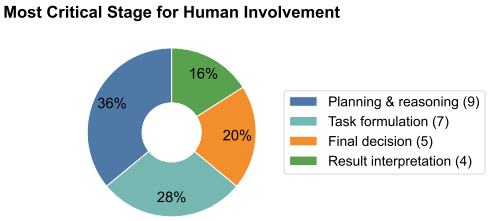

Figure 5: Most critical stages for human involvement.

assistant, and only 12% expect fully autonomous researchers (Figure 6).

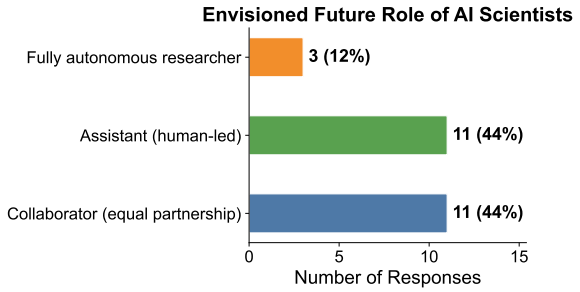

Figure 6: Envisioned future roles of AI scientists.

E.3.4 Perceived Issues and Severity

Figure 7 shows Safety & Reliability is rated most severe (mean=3.16/5), with 40% considering it "Very" or "Extremely" problematic.

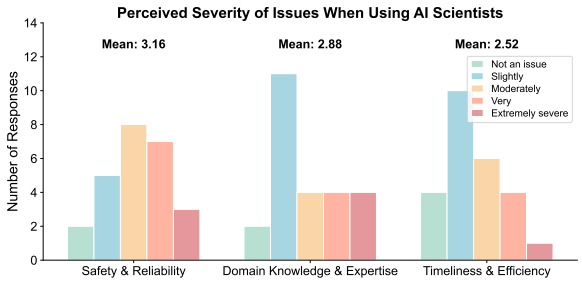

Figure 7: Perceived severity of issues when using AI scientists.

E.3.5 Superhuman-Level Performance Timeline

Figure 8 shows **64% expect AI to reach superhuman-level research performance within 5 years**. Only 4% consider it impossible.

E.4 Open-Ended Responses

E.4.1 Biggest Limitations (Q10)

Table 2 categorizes responses on AI scientists' biggest limitations. **Hallucination** (28%) and **lack of creativity** (24%) dominate.

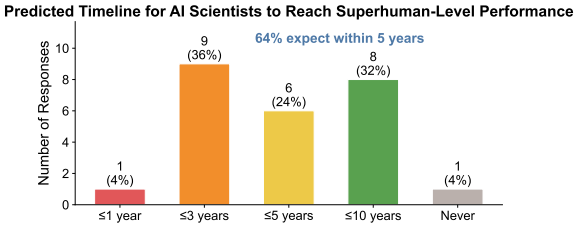

Figure 8: Predicted timeline for superhuman-level AI research performance.

E.4.2 Future Concerns (Q12)

Table 3 presents concerns about AI scientists' future. **Misinformation** (25%) and **academic quality degradation** (12.5%) are prominent.

E.5 Summary of Key Findings

1. **High Adoption:** 80% frequently use AI for research; coding (100%), writing (80%), and literature review (76%) are dominant use cases.
2. **Execution-Innovation Gap:** Execution tasks rated Good (coding: 3.92, writing: 3.68) while creative tasks rated Poor (hypothesis: 2.71, reproduction: 2.63).
3. **Hallucination as Primary Concern:** 28% cite hallucination as the biggest limitation; 25% worry about future misinformation.
4. **Human Oversight Critical:** 64% believe early stages require human involvement; only 16% accept minimal intervention.
5. **Optimistic but Cautious:** 64% expect superhuman performance within 5 years, but 88% prefer human-led or collaborative models over full autonomy.

E.6 Limitations

The sample (N=25) is relatively small and concentrated in Computer Science (96%), limiting generalizability. Participants are active AI users, potentially overestimating adoption. Findings reflect January 2026 perspectives.

E.7 Full Survey Questionnaire

Table 2: Reported limitations of current AI scientists (N=25), categorized by theme.

Category	**N (%)**	**Representative Responses**
Hallucination / Factual Accuracy	7 (28%)	"Hallucination everywhere"; "When we ask to find related papers, the title often looks reasonable but it actually returns paper that doesn't exist"; "Over confident on things that it is not sure"
Creativity / Innovation	6 (24%)	"They are horrible at ideation and experimental design"; "Limited ability to generate truly original ideas"; "Ideas generated are sometimes naive"; "Incremental idea without novelty"
Domain Knowledge	4 (16%)	"Context and understanding abilities"; "Do not know enough"; "Multi-modal ability"
Reliability / Accountability	3 (12%)	"Results in papers/model cards do not match real-world performance"; "Accountability in problem formulation"
Speed / Efficiency	2 (8%)	"Time cost"; "Speed is too slow"
Other	3 (12%)	"No reliable long-term memory"; "Need deep thinking capability"; "Access to data is crucial"

Table 3: Concerns about the future of AI scientists (N=24), categorized by theme.

Category	**N (%)**	**Representative Responses**
Hallucination / Misinformation	6 (25%)	"Generating fake facts"; "Producing research from hallucinated results—we don't understand why AI proposed the method"; "Ability to verify information"
Research Quality / Academic Impact	3 (12.5%)	"Watery papers beat potential good papers"; "Increased submissions result in lack of reviewers and poor review quality"
Safety / Reliability	3 (12.5%)	"Errors matter significantly in science"; "Reliability and performance"
Job Displacement / Human Role	3 (12.5%)	"Replace jobs"; "Humans stop thinking $\Rightarrow$ new knowledge disappears"
Creativity / Novelty	2 (8.3%)	"AI won't have creativity needed for fully autonomous research"; "Lack of good taste for novelty"
Capability / Cost	5 (20.8%)	"Will they be smart enough?"; "Cost"; "Data access limitations"
Other	2 (8.3%)	"Accuracy in complex research"; "Not reliable"

AI Scientists Survey

This survey aims to understand how researchers perceive, use, and envision the future of AI Scientists. All responses are anonymous and will be used for academic research purposes only.

Section 1: Demographics

Q1.* What is your primary research domain?

☐ Computer Science ☐ Physics/Chemistry/Biology ☐ Social Science ☐ Other

Q2.* What is your highest level of education?

☐ Undergraduate ☐ Master's ☐ PhD ☐ Postdoc ☐ Faculty

Q3.* How many years of research experience do you have?

☐ <1 year ☐ 1–3 years ☐ 3–5 years ☐ 5–10 years ☐ >10 years

Section 2: Current Usage

Q4.* Have you used AI-based systems (LLMs, agents, automated pipelines) to assist with research tasks?

☐ Frequently ☐ Occasionally ☐ Tried a few times ☐ Never

Q5.* In which research scenarios do you use AI scientists? (Select all that apply)

☐ Paper writing ☐ Literature review ☐ Coding/debugging ☐ Experiment design
☐ Data analysis ☐ Hypothesis generation ☐ Tool usage ☐ Reproducing papers

Q5-b.* Rate AI capability in each module (1=Very poor to 5=Excellent):

Paper writing | Literature review | Coding | Experiment design | Data analysis | Hypothesis generation | Tool usage | Reproducing papers

Q6.* What types of AI systems do you mainly use? (Select all that apply)

☐ General-purpose LLMs (ChatGPT, Claude, Gemini) ☐ Domain-specific AI tools ☐ Autonomous research agents
☐ Coding agents (Cursor, Copilot) ☐ Open-source frameworks (Sakana AI Scientist) ☐ Deep research

Section 3: Human-AI Collaboration

Q7.* How are you involved when using AI scientists? (Select all that apply)

☐ Prompting/instruction design ☐ Intermediate verification ☐ Providing feedback ☐ Final decision by human ☐ Minimal intervention

Q8.* At which stage is human involvement MOST critical?

☐ Task formulation/goal setting ☐ Planning and reasoning ☐ Tool execution ☐ Result interpretation ☐ Final decision

Q9.* Rate severity of issues (1=Not an issue to 5=Extremely severe):

Safety & Reliability | Timeliness & Efficiency | Domain Knowledge & Expertise

Section 4: Future Perspectives

Q10.* What is the single biggest limitation of current AI scientists? *(Open-ended)*

Q11.* How do you envision AI scientists working with humans in the future?

☐ Assistant (human-led) ☐ Collaborator (equal partnership) ☐ Supervisor (AI-led) ☐ Fully autonomous ☐ Not sure

Q12.* What concerns you MOST about the future of AI scientists? *(Open-ended)*

Q13.* When will AI scientists reach superhuman-level performance in your field?

☐ Within 1 year ☐ Within 3 years ☐ Within 5 years ☐ Within 10 years ☐ Never/Not possible

Q14. Any additional comments or suggestions? *(Optional)*

Q15.* Do you agree to allow your responses to be used for academic research? ☐ Yes, I agree ☐ No

Figure 9: Complete survey questionnaire. Questions marked with * were required.