Agent Systems for Academic Research Automation

Pierfrancesco Beneventano [1] **Riccardo Neumarker** [2 1] **Theodoros Evgeniou** [3] **Mahmoud Abdelmoneum** [1] **Marc Bacvanski** [1] **Yulu Gan** [1] **Mehdi Hajoub** [1] **Qianli Liao** [1] **Emanuele Rimoldi** [1 4] **Kushagra Tiwary** [1] **Andrea Pinto** [5] **Liu Ziyin** [1 6] **Tomer Galanti** [7] **Tomaso Poggio** [1]

Abstract

AI scientist systems increasingly propose hypotheses, plan experiments, execute analyses, draft papers, and participate in review-like workflows. As these systems move from local assistance toward producing claim-bearing scientific artifacts, aggregate descriptors such as agent count, tool use, or autonomy no longer specify what a system contributes, what artifact it produces, or what can verify the claims it helps make. We introduce *verifier-matched autonomy*: the principle that a system's apparent autonomy is meaningful only relative to the strength, cost, and latency of the verifier available for those claims. We develop a three-dimensional framework organized by research-lifecycle coverage, artifact type, and verification regime, and use it to position citation-grounded writing systems, survey generators, full-paper research pipelines, review agents, and domain-specific discovery systems. This framework treats labels such as tool, co-author, and founder as workflow-specific roles rather than system-wide properties. We derive implications for evaluation, provenance, and governance, emphasizing *closure failure*: the risk that claim-bearing scientific artifacts appear settled before their underlying claims have been adequately verified.

1. Introduction

AI scientist systems now span a continuum from local assistance to increasingly autonomous participation in scientific workflows. They retrieve and synthesize literature, draft scholarly text, generate reviews, design experiments, execute code, analyze results, and in some cases produce complete research manuscripts (Wang et al., 2024; Liang et al., 2025; Asai et al., 2024; Ifargan et al., 2025; Lu et al., 2024; Schmidgall et al., 2025; Jin et al., 2024; Chamoun et al., 2024). This expansion creates a conceptual problem: systems are often described as tools, co-authors, founders, agents, or autonomous scientists, but those labels do not specify what the system contributes, what scholarly artifact it produces, or what can verify the claims that artifact makes.

This distinction matters because scientific artifacts are not merely outputs. They are the objects through which claims become visible to reviewers, institutions, authors, and downstream readers. A system that proposes a hypothesis, writes a section, runs a benchmark, or drafts a rebuttal is participating in different parts of the scientific process and should be judged under different obligations. A citation-grounded paragraph, a literature survey, a full paper built from executed analyses, and a peer review can all be produced by language-model agents, but they are not answerable to the same kind of evidence.

We focus on the *claim-bearing layer* of AI scientist systems: the point at which automated workflows become scholarly artifacts such as sections, surveys, full papers, reports, reviews, rebuttals, protocols, or structured revisions. We do not claim that AI scientist systems are only manuscript systems. Rather, we argue that this layer is central because it is where automated scientific work is converted into explicit claims, credit, evaluation, and institutional decisions. If this conversion is unreliable, an AI scientist can produce a document that appears scientifically complete even when its underlying claims remain only partially checked.

The central risk is therefore not only local hallucination. It is *closure failure*: the production of a scientific artifact that linguistically resolves uncertainty before the underlying science has been adequately resolved. Closure failure can occur even when individual sentences are fluent, citations are present, code executes, or a review sounds plausible. What matters is whether the strength of the artifact's claims is matched to the strongest verifier available in the workflow. By a *verifier*, we mean any mechanism that can materially test or constrain a claim: citation support, executable code and data, a proof assistant, simulator or instrument feedback,

[1]Massachusetts Institute of Technology [2]ETH Zürich [3]INSEAD [4]EPFL [5]Notte [6]NTT Research [7]Texas A&M University. Correspondence to: Pierfrancesco Beneventano <pierb@mit.edu>.

Accepted by the AI for Science Workshop at ICML 2026.

or external empirical validation.

We call autonomy *verifier-matched* when a system's degree of initiative and the strength of its claims are aligned with the verifier available for those claims. This framing shifts attention away from agent count, tool use, or autonomy as a scalar property. A system can be highly agentic yet weakly verified, or narrow in scope yet strongly checked. Conversely, a system may be founder-like in proposing directions, co-author-like in drafting and revising artifacts, and tool-like with respect to validation that still depends on humans, instruments, or external experiments.

Relative to broader surveys of autonomous scientific discovery and research agents (Wei et al., 2025; Zhang et al., 2025), our goal is not to provide another exhaustive inventory of systems. We instead isolate the claim-bearing layer of AI scientist workflows and ask when the autonomy expressed in a scholarly artifact is warranted by the verifier available for its claims. This shifts the comparison from system labels and agent architectures to artifact obligations, claim types, and verification bottlenecks.

Contributions. We make four contributions. First, we define a claim-bearing scope for AI scientist systems: the layer where automated workflows become scholarly artifacts that affect review, credit, and institutional decisions. Second, we introduce and instantiate a three-dimensional framework based on task coverage, artifact type, and verification regime, showing how representative system families differ in artifact obligations and verifier bottlenecks. Third, we synthesize the principle of verifier-matched autonomy, showing why autonomy depends on the strength, cost, and latency of the available verifier rather than on agent count alone. Fourth, we derive implications for evaluation, provenance, and governance, including why tool, co-author, and founder should be assigned to workflow roles rather than systems as wholes.

2. Scope: The Claim-Bearing Layer

We include systems whose outputs directly shape claim-bearing scholarly artifacts: citation-grounded passages, literature reviews, full papers, research reports, peer reviews, rebuttals, revision plans, protocols, or manuscript-state edits. This scope is narrower than AI for science as a whole, but it is not merely about writing style. It concerns the interface where scientific work is represented as claims that others may cite, review, reproduce, fund, or trust.

The scope also separates two questions that are often conflated. One question is whether an AI system can produce or execute scientific work. Another is whether the resulting artifact states only what the evidence can justify. A laboratory controller, molecule generator, or benchmark optimizer may be scientifically important even when manuscript production is not its primary output. Such systems become central to our analysis when their outputs are used to support claim-bearing artifacts, because the relevant question then becomes how the artifact's claims are grounded, checked, and limited.

This distinction helps handle boundary cases. Systems that automate experiments but do not produce scholarly artifacts are not core examples of the claim-bearing layer, although they provide evidence about verification regimes. Systems that write from completed research materials belong inside the scope, but should be distinguished from systems that also generate and validate the underlying research process. Systems with limited public documentation may still be important, but they support weaker claims about architecture and reliability than systems with peer-reviewed papers, technical artifacts, or reproducible repositories.

Evidence tiers and boundary cases. Because AI scientist systems are documented through heterogeneous sources, we distinguish the visibility of a system from the strength of evidence available about it. Peer-reviewed papers and reproducible technical artifacts support stronger architectural and reliability claims than public demos, official announcements, or informal repositories. Boundary cases are therefore included when they clarify the claim-bearing layer, but not treated as equally evidential. For example, systems whose primary output is an experimental loop, a code patch, or a public workflow demonstration may be informative about verification regimes without being core examples of claim-bearing AI scientist systems.

Our comparison therefore asks three questions of each system. Where does it intervene in the research lifecycle? What artifact is it answerable for producing or revising? What verifier governs the principal claims made by that artifact? The answer to these questions determines not only where a system fits in the landscape, but also what kinds of autonomy and institutional responsibility are warranted.

3. A Three-Dimensional Map

A useful framework for AI scientist systems should do more than name examples. It should help position a newly introduced system before it is absorbed into a retrospective catalog, and it should make clear which comparisons are meaningful. We therefore organize the landscape along three dimensions: task coverage, artifact type, and verification regime. No single dimension is sufficient on its own. A system's location in the research lifecycle does not determine what artifact it produces, and artifact type does not determine what can verify its claims. Figure 1 summarizes the framework.

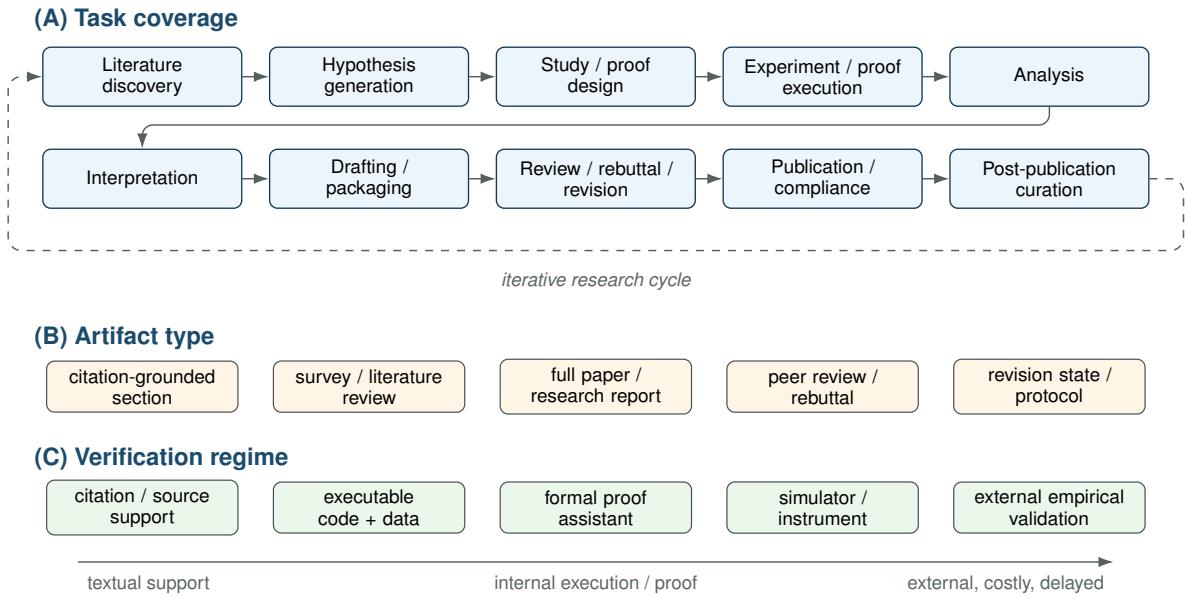

Figure 1. A three-dimensional map for claim-bearing AI scientist systems. Panel A adapts the research lifecycle into stages at which systems may intervene. Panel B lists the scholarly artifacts for which such systems may be answerable. Panel C lists representative verification regimes; these are not a single quality ranking, but they determine what kind of autonomy and claim strength are warranted.

3.1. Task Coverage

Task coverage, shown in Figure 1A, describes where a system acts within the research lifecycle. Some systems operate narrowly, for example by drafting citation-grounded passages or assisting with related-work generation (Wang et al., 2025a). Survey systems occupy a broader literature-to-manuscript region: they construct corpora, form outlines, synthesize sources, and generate long-form reviews (Wang et al., 2024; Yan et al., 2025; Liang et al., 2025; Chen et al., 2025a; Wang et al., 2025c). Full-paper systems extend further into hypothesis generation, experiment design, execution, analysis, and manuscript production (Ifargan et al., 2025; Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Tang et al., 2025). Review and rebuttal systems operate later in the cycle, where the relevant state is an existing artifact together with reviewer comments, criticism, and revision history (Jin et al., 2024; Gao et al., 2025a; Chamoun et al., 2024).

Task coverage should not be read as a simple autonomy scale. A review agent is not merely a weaker version of an end-to-end research pipeline; it is answerable to a different object and fails in different ways. Likewise, a system that automates experiments but does not write or revise claim-bearing artifacts should not be evaluated by the same criteria as a system that produces a paper from retrieved literature. The relevant question is not whether a system covers "more" of the task in the abstract, but where its decisive work occurs and what later stages need in order to trust or revise that work.

3.2. Artifact Type

Artifact type, shown in Figure 1B, captures the scholarly object a system is responsible for producing or revising. We distinguish four recurrent families in the current literature. Section-level systems are answerable for local claim–citation alignment. Survey and literature-review systems are answerable for corpus selection, synthesis, disagreement representation, and document-level coherence. Full-paper systems are answerable for the relationship between manuscript claims and an underlying research process. Review, revision, and rebuttal systems are answerable for critique, response, and improvement of an existing artifact.

Artifact type matters because each object imposes a different contract. A section writer can appear successful if each local claim is supported by an appropriate source, but that criterion is too weak for a survey whose value depends on coverage, synthesis, and the representation of disagreement. A full-paper agent may produce a rhetorically coherent manuscript while overstating the experimental evidence behind it. A review agent may generate comments that are useful and plausible without actually detecting scientific error. These differences are obscured when systems are compared only by model, agent count, or output length.

3.3. Verification Regime

Verification regime, shown in Figure 1C, describes what can check the principal claims made by a system's artifact. In formally verifiable symbolic domains, proof assistants can provide exact feedback once claims are formalized (Feng

et al., 2026; Ospanov et al., 2025). In executable code-and-data domains, claims can often be checked through rerunnable analyses, benchmark harnesses, and provenance trails (Ifargan et al., 2025; Lu et al., 2024; Schmidgall et al., 2025). In tool-mediated physical sciences, simulators and instruments provide strong but still proxy validation (Boiko et al., 2023; Szymanski et al., 2023; Hill & Ryoo, 2026). In open empirical biological and clinical domains, decisive validation may remain external, slow, costly, and ethically constrained (Google Research & Google DeepMind, 2025; Roohani et al., 2025; Wang et al., 2025b).

This dimension determines what kind of autonomy is warranted. When the verifier is exact, internal, and cheap to invoke, more initiative can safely move into the agent loop. When the verifier is external, delayed, or only indirectly related to the artifact's central claims, the system must be more conservative: it can triage hypotheses, rank experiments, draft protocols, or summarize evidence, but should not write as though validation has already closed. Similar task coverage can therefore require different architectures depending on the verifier. Conversely, systems that look different on the surface may share the same design pressure if their claims are governed by the same validation bottleneck.

3.4. Using the Map as a Diagnostic

The framework is useful only if it changes how systems are interpreted. Table 1 illustrates this diagnostic use on representative systems and system families. The point is not to provide an exhaustive catalog, but to separate questions that are often collapsed: what artifact the system produces, what can check it, what autonomy is warranted, and what characteristic failure appears if the system overclaims.

The examples draw on citation-grounded writing systems (Wang et al., 2025a), survey generators (Wang et al., 2024; Liang et al., 2025; Yan et al., 2025), full-paper research pipelines (Ifargan et al., 2025; Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025), review and rebuttal agents (Jin et al., 2024; Chamoun et al., 2024), formal proof agents (Feng et al., 2026; Ospanov et al., 2025), tool-mediated laboratory or simulation systems (Boiko et al., 2023; Szymanski et al., 2023; Hill & Ryoo, 2026), and biomedical proposal systems (Google Research & Google DeepMind, 2025; Roohani et al., 2025). These examples differ in evidence tier; systems with lower public documentation are used diagnostically to clarify boundary cases, not as equal-strength empirical baselines.

Table 1. Compact diagnostic readings under the three-dimensional framework. Examples differ in evidence tier; roles are assigned to artifact–verifier pairs, not to systems as wholes.

System or family	Artifact and verifier / proxy	Diagnostic reading
SCHOLARCOPILOT	Section or passage; retrieved sources and claim–citation checks	Local co-author for drafting; risk that citation presence masks weak evidential adequacy.
AUTOSURVEY, SURVEYX, SURVEYFORGE	Survey; corpus construction, source support, and structured synthesis checks	Synthesis co-author; risk of coverage gaps, smoothed disagreement, or weak research guidance.
DATA-TO-PAPER, THE AI SCIENTIST, AGENT LABORATORY	Full paper or report; execution traces, logs, benchmarks, and provenance where available	Stronger autonomy for traceable claims; risk that narrative confidence outruns executed evidence.
AGENTREVIEW, SWIF^{2}T	Review or rebuttal; review-quality proxies such as human agreement, usefulness, and actionability	Reviewer assistant; risk that plausible critique is mistaken for scientific verification.
Formal proof agents	Formal claim or proof; proof-assistant feedback after formalization	High autonomy after formalization; risk of plausible prose proofs without formal validity.
COSCIENTIST, A-LAB, GRACE	Protocol, experiment, or simulation output; simulator, instrument, or characterization loop	Locally autonomous execution; risk that proxy success is mistaken for final physical truth.
AI CO-SCIENTIST, BIODISCOVERYAGENT	Hypothesis or ranked experiment; literature, existing data, and later empirical validation	Founder-like triage before validation; risk that speculation is presented as validated discovery.

4. Verifier-Matched Design Lessons

The diagnostic readings in Table 1 expose recurring design constraints. The architectures that support stronger scientific claims are not simply those with more agents, more tools, or longer context windows. They are systems whose artifact obligations are aligned with the kind of evidence and verification their architecture can actually provide.

4.1. Claims Should Not Outrun Their Verifiers

Different scientific claims require different forms of support. Literature-backed claims can be constrained through retrieval, citation selection, and support checking. Empirical performance claims require executable analysis, rerunnable code, and provenance. Formal claims require proof checking. Causal, biological, or clinical claims may require external experiments that cannot be closed within the agent loop.

The core design principle is therefore claim-to-verifier matching. A system should not make stronger claims than its strongest available verifier can justify. This is why retrieval alone is too weak for empirical performance claims, why code execution is too weak for biological mechanisms requiring external validation, and why peer-review-like critique is too weak for establishing factual correctness. When this matching fails, the system may still produce a fluent and locally plausible artifact, but the artifact can present partial evidence as settled science. This is closure failure at the document level.

4.2. Project State Is an Accountability Object

Long-horizon scholarly artifacts require more than independent generations over a large context window. They depend on information that must remain available, revisable, and contestable across many steps: outlines, source selections, methodological decisions, experimental outcomes, caveats, reviewer comments, and document edits. Survey systems increasingly maintain structured representations of source material and evolving plans, for example through memory-driven retrieval, attribute trees, or memory-guided writing (Yan et al., 2025; Liang et al., 2025; Chen et al., 2025a). Full-paper and manuscript-state systems similarly benefit from explicit provenance, execution history, and revision state, as in systems that trace numerical claims to analysis code or track live document edits and reviewer-driven revisions (Ifargan et al., 2025; Hou et al., 2025).

This is not only a capability issue; it is an accountability issue. If a limitation discovered late in the pipeline should narrow the abstract, or if a failed experiment should weaken the discussion, the system needs a representation of the dependency between evidence and claim. A larger context window may make the whole draft visible, but it does not by itself encode which claims depend on which sources, analyses, or reviewer decisions. Claim-bearing AI scientist systems therefore need explicit project state: records of claims, evidence, uncertainty, and revision dependencies that can be inspected and updated.

4.3. Critique Is Not Verification

Review and revision agents can improve specificity, coverage, actionability, and clarity (Jin et al., 2024; Gao et al., 2025a; Zhu et al., 2025; Chamoun et al., 2024). Similar critique loops appear inside survey and full-paper systems, where reviewer-like agents or rubrics are used to refine drafts. These loops are useful, but they should not be mistaken for verification.

Critique usually operates over representations already available to the system: the draft, the retrieved sources, the rubric, or an internal deliberation trace. It can identify that a claim seems unsupported, poorly framed, or inconsistent with another section. It cannot by itself establish whether an experiment was correctly executed, whether a statistical analysis is valid, whether a cited source truly supports the claim, or whether an external biological mechanism has been validated. Those checks require contact with something outside the text: data, code, formal proof states, instruments, or empirical outcomes.

The failure mode is self-confirming criticism. A system may iteratively polish a manuscript, add caveats, and generate plausible reviews while leaving the truth conditions of the central claims unchanged. Evaluation should therefore distinguish critique that improves perceived quality from verification that increases epistemic reliability.

4.4. Autonomy Is Coupled to Verifier Strength, Cost, and Latency

Autonomy should not be treated as a scalar property of a system. It is a relation between initiative and verification. In formal domains, once a claim is formalized, the proof assistant can close parts of the loop internally. In executable code-and-data domains, agents can rerun analyses, compare outputs, and trace numerical claims back to code. In physical sciences, simulators and instruments can close local loops, but tool success may still be only a proxy for the final scientific claim. In open empirical biology and medicine, decisive validation may require wet-lab experiments, patient cohorts, or clinical procedures outside the agent loop.

This produces a graded view of tool, co-author, and founder-like behavior. A system may be founder-like when proposing research directions, co-author-like when drafting and revising artifacts, and tool-like when validating claims that still require human oversight or external experiments. The same system can therefore occupy different roles within one workflow. What matters is not the global label assigned to the system, but the verifier available for each class of claim it helps produce.

5. Evaluation: From Local Scores to Claim-Level Validity

Current evaluations of AI scientist systems remain fragmented because they test different artifacts under different validators. Citation-grounding benchmarks evaluate whether generated text is locally supported by sources (Gao et al., 2023; Funkquist et al., 2023; Ravichander et al., 2025). Survey benchmarks move toward long-form synthesis, but still struggle to capture whether the right literature was selected, whether disagreement was represented, or whether the document offers useful research guidance (Sun et al., 2025). Execution-backed benchmarks test whether agents can reproduce analyses, solve scientific tasks, or operate under external checking (Starace et al., 2025; Chen et al., 2025b; OpenAI, 2025). Review benchmarks test plausibility, usefulness, and agreement with human review patterns (Gao et al., 2025b; Liang et al., 2024; Thakkar et al., 2025).

These evaluations are valuable, but they do not measure a single underlying quantity called "AI scientist quality." A strong result on citation grounding is not evidence that a system can run valid experiments. A strong result on code execution is not evidence that the resulting paper states its contribution honestly. A useful generated review is not evidence that the review detects consequential scientific errors. The evaluation object changes with the artifact and

the verifier.

For claim-bearing systems, a manuscript alone is often too thin to evaluate, because it hides how claims entered the document and what supports them. The relevant evaluation object is therefore not just an artifact, but an artifact together with its evidential state.

A minimal claim-level evaluation object. A verifier-matched benchmark should evaluate not only an output artifact, but the artifact together with its evidential state. A minimal evaluation bundle should include: (i) extracted central claims, (ii) the evidence object attached to each claim, such as sources, code, data, proof states, instrument outputs, or external validation records, (iii) the verifier type used to assess the claim, and (iv) revision history showing whether later edits preserved, weakened, or overstated the evidential basis. This bundle makes it possible to evaluate claim discipline: whether the artifact states only what its evidence can support.

Three benchmark gaps follow. First, current benchmarks rarely test global artifact coherence: whether the abstract, methods, results, discussion, and limitations remain scientifically consistent as evidence changes. Second, they rarely test contradiction handling across sources, even though scientific synthesis often depends on representing disagreement rather than smoothing it away. Third, they rarely test validator–claim matching: whether the mechanism used to score an output is strong enough for the kind of claim being made. Without this third check, systems can appear well evaluated while their central claims remain under-verified.

Stress tests for closure failure. Evaluation should also perturb the evidential state and test whether the artifact updates correctly. If a cited source is removed, a failed experiment is introduced, a proof state fails, a simulator objective is revealed to be a proxy, or a reviewer identifies a limitation, the system should propagate the change to the abstract, contribution statement, discussion, and limitations. Systems that preserve fluent conclusions while their evidence weakens are exhibiting document-level closure failure.

6. Governance: Roles, Responsibility, and Closure Failure

The same framework changes how governance should be framed. The categories "tool," "co-author," and "founder" should not be assigned to a system as a whole. They should be assigned to roles within a scientific workflow and, more precisely, to classes of claims within that workflow. A system may act as a tool when retrieving sources, executing code, or performing local checks; as a co-author-like participant when synthesizing literature, framing contributions, drafting, interpreting, or revising artifacts; and as a founder-like participant when proposing directions, prioritizing hypotheses, or allocating experimental attention. Each role creates different obligations because each role produces different claims under different verifiers.

A minimal disclosure unit. A practical disclosure unit is not merely "AI used," but a role–claim–verifier–uncertainty tuple: which role the system played, which claims or artifact components it shaped, what verifier was available for those claims, and what uncertainty remained unresolved. This tuple is small enough to be reported in author disclosures or artifact cards, but more informative than a binary AI-use statement.

This role-based view also clarifies responsibility. If an AI scientist contributes only text polish, ordinary disclosure may be sufficient. If it selects literature, frames the contribution, proposes hypotheses, interprets results, or drafts responses to reviewers, then the relevant question is not merely whether a tool was used, but which claims were shaped by the system and how those claims can be inspected. Responsibility becomes difficult to exercise when authors cannot reconstruct how an idea, citation, caveat, comparison, or inferred limitation entered the artifact. For this reason, governance should track role, claim, verifier, and uncertainty together rather than treating AI use as a single binary disclosure.

Provenance is therefore more than a record-keeping preference. It is a condition for contestability. In ordinary scientific practice, reproducibility attaches to data, code, protocols, and analysis pipelines. In claim-bearing AI scientist systems, important epistemic work also occurs in retrieval states, prompts, model versions, memory stores, tool traces, and revision loops. A paper may be experimentally reproducible while remaining procedurally opaque as a generated artifact. Conversely, a generation trace may be replayable without showing that the claims it produced were warranted.

Closure failure is the governance risk that connects these points. When a system writes with confidence before validation has closed, institutions may evaluate a finished-looking artifact rather than the evidence state behind it. This risk is especially acute if authors use agents to optimize for reviewer-legible claims while reviewers use similar systems to generate critique. The result could be a closed loop of mutually fluent but insufficiently grounded scientific evaluation. Governance should therefore track not only whether AI was used, but what role it played, what claims resulted, what verifier was available, and where uncertainty remained unresolved.

7. Conclusion

AI scientist systems should not be compared only by agent count, tool use, or apparent autonomy. The more important question is what claim-bearing artifact a system produces, which research tasks it performs, and what can verify its principal claims. This three-dimensional view clarifies why superficially similar systems may require different architectures and why systems with similar autonomy claims may deserve different levels of trust.

The central lesson is verifier-matched autonomy: initiative should expand only where the corresponding verifier can constrain the resulting claims. AI scientist systems become reliable only when their initiative and their claims remain aligned with the strongest verifier available in the workflow. Future systems will need not only better models and more tools, but richer representations of claim status, evidence, uncertainty, provenance, and revision dependencies. Without those structures, faster and more fluent AI scientists may simply produce more convincing artifacts whose claims have not been adequately established.

Impact Statement

This paper analyzes AI systems that participate in scientific writing, review, and research automation. Potential benefits include faster synthesis of scientific evidence, broader access to research assistance, and improved support for reproducibility when systems expose provenance and claim traces. Potential risks include overclaiming, weakened attribution, unreliable review loops, premature trust in generated artifacts, and institutional confusion about responsibility. We argue that verifier-matched autonomy, provenance, and claim-level evaluation are necessary safeguards for reducing these risks.

References

Asai, A., He, J., Shao, R., Shi, W., Singh, A., Chang, J. C., Lo, K., Soldaini, L., Feldman, S., D'arcy, M., Wadden, D., Latzke, M., Tian, M., Ji, P., Liu, S., Tong, H., Wu, B., Xiong, Y., Zettlemoyer, L., Neubig, G., Weld, D., Downey, D., tau Yih, W., Koh, P. W., and Hajishirzi, H. Openscholar: Synthesizing scientific literature with retrieval-augmented lms, 2024. URL https://arxiv.org/abs/2411.14199.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0.

Chamoun, E., Schlichtkrull, M., and Vlachos, A. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://aclanthology.org/2024.findings-acl.580/. Findings of ACL 2024. arXiv:2405.20477.

Chen, J., Yang, Z., Shen, Y., Liu, J., Belloum, A., Grosso, P., and Papagianni, C. SurveyGen-I: Consistent scientific survey generation with evolving plans and memory-guided writing. *arXiv preprint arXiv:2508.14317*, 2025a. doi: 10.48550/arXiv.2508.14317. URL https://arxiv.org/abs/2508.14317.

Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., Dey, V., Xue, M., Baker, F. N., Burns, B., Adu-Ampratwum, D., Huang, X., Ning, X., Gao, S., Su, Y., and Sun, H. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery, 2025b. URL https://arxiv.org/abs/2410.05080.

Feng, T., Trinh, T. H., Bingham, G., Hwang, D., Chervonyi, Y., Jung, J., Lee, J., Pagano, C., hyun Kim, S., Pasqualotto, F., Gukov, S., Lee, J. N., Kim, J., Hou, K., Ghiasi, G., Tay, Y., Li, Y., Kuang, C., Liu, Y., Lin, H., Liu, E. Z., Nayakanti, N., Yang, X., Cheng, H.-T., Hassabis, D., Kavukcuoglu, K., Le, Q. V., and Luong, T. Towards autonomous mathematics research. *arXiv preprint arXiv:2602.10177*, 2026. doi: 10.48550/arXiv.2602.10177.

Funkquist, M., Kuznetsov, I., Hou, Y., and Gurevych, I. CiteBench: A benchmark for scientific citation text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2212.09577. EMNLP 2023. arXiv:2212.09577.

Gao, T., Yen, H., Yu, J., and Chen, D. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2305.14627. EMNLP 2023. arXiv:2305.14627.

Gao, X., Ruan, J., Gao, J., Liu, T., and Fu, Y. ReviewAgents: Bridging the gap between human and AI-generated paper reviews. *arXiv preprint arXiv:2503.08506*, 2025a. doi: 10.48550/arXiv.2503.08506. URL https://arxiv.org/abs/2503.08506.

Gao, X., Ruan, J., Zhang, Z., Liang, Q., Wu, G., Song, S., Liu, B., and Sun, C. MMReview: A multidisciplinary and multimodal benchmark for LLM-based peer review automation, 2025b. URL https://arxiv.org/abs/2508.14146. arXiv:2508.14146.

Google Research and Google DeepMind. Accelerating scientific breakthroughs with an AI co-scientist. https://research.google/blog/accelerating-scientific-breakthroughs-with-ai-co-scientist/, 2025. Official research blog post.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomasev, N., Guan, Y., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penadés, J. R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A., and Natarajan, V. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025. doi: 10.48550/arXiv.2502.18864. URL https://arxiv.org/abs/2502.18864.

Hill, J. and Ryoo, H. J. Grace: an agentic ai for particle physics experiment design and simulation, 2026. URL https://arxiv.org/abs/2602.15039.

Hou, J., Huikai, A. L., Chen, N., Gong, Y., and He, B. Paperdebugger: A plugin-based multi-agent system for in-editor academic writing, review, and editing. *arXiv preprint arXiv:2512.02589*, 2025. URL https://arxiv.org/abs/2512.02589.

Ifargan, T., Hafner, L., Kern, M., Alcalay, O., and Kishony, R. Autonomous LLM-driven research — from data to human-verifiable research papers. *NEJM AI*, 2(1), 2025. doi: 10.1056/AIoa2400555. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400555.

Jin, Y., Zhao, Q., Wang, Y., Chen, H., Zhu, K., Xiao, Y., and Wang, J. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://arxiv.org/abs/2406.12708. EMNLP 2024, Oral. arXiv:2406.12708. Repository: https://github.com/ahren09/agentreview.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., and Zou, J. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 2024. URL https://arxiv.org/abs/2310.01783. arXiv:2310.01783.

Liang, X., Yang, J., Wang, Y., Tang, C., Zheng, Z., Song, S., Lin, Z., Yang, Y., Niu, S., Wang, H., Tang, B., Xiong, F., Mao, K., and Li, Z. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025. URL https://arxiv.org/abs/2502.14776.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. doi: 10.48550/arXiv.2408.06292. URL https://arxiv.org/abs/2408.06292.

OpenAI. FrontierScience: Evaluating AI's ability to perform scientific research tasks. OpenAI Technical Report, December 2025. URL https://openai.com/index/frontierscience/. Full paper available at https://cdn.openai.com/pdf/2fcd284c-b468-4c21-8ee0-7a783933efcc/frontierscience-paper.pdf.

Ospanov, A., Farnia, F., and Yousefzadeh, R. Apollo: Automated llm and lean collaboration for advanced formal reasoning, 2025. URL https://arxiv.org/abs/2505.05758.

Ravichander, A., Ghela, S., Wadden, D., and Choi, Y. HALoGEN: Fantastic LLM hallucinations and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL https://arxiv.org/abs/2501.08292. ACL 2025 Long Papers. arXiv:2501.08292. Anthology: https://aclanthology.org/2025.acl-long.71/.

Roohani, Y., Lee, A., Huang, Q., Vora, J., Steinhart, Z., Huang, K., Marson, A., Liang, P., and Leskovec, J. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments, 2025. URL https://arxiv.org/abs/2405.17631.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., and Barsoum, E. Agent laboratory: Using LLM agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025. doi: 10.48550/arXiv.2501.04227. URL https://arxiv.org/abs/2501.04227.

Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., Heidecke, J., Glaese, A., and Patwardhan, T. Paperbench: Evaluating ai's ability to replicate ai research, 2025. URL https://arxiv.org/abs/2504.01848.

Sun, Z., Zhu, X., Zhou, X., Tong, X., Wang, S., Fu, J., Li, G., Liu, Z., and Wu, F. SurveyBench: Can LLM(-agents) write academic surveys that align with reader needs? *arXiv preprint arXiv:2510.03120*, 2025. doi: 10.48550/arXiv.2510.03120. URL https://arxiv.org/abs/2510.03120.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D.,

Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., and Ceder, G. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06734-w.

Tang, J., Xia, L., Li, Z., and Huang, C. AI-researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025. doi: 10.48550/arXiv.2505.18705. URL https://arxiv.org/abs/2505.18705.

Thakkar, N., Yuksekgonul, M., Silberg, J., Garg, A., Peng, N., Sha, F., Yu, R., Vondrick, C., and Zou, J. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025, 2025. URL https://arxiv.org/abs/2504.09737.

Wang, Y., Guo, Q., Yao, W., Zhang, H., Zhang, X., Wu, Z., Zhang, M., Dai, X., Zhang, M., Wen, Q., Ye, W., Zhang, S., and Zhang, Y. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*, 2024. doi: 10.48550/arXiv.2406.10252. URL https://arxiv.org/abs/2406.10252.

Wang, Y., Ma, X., Nie, P., Zeng, H., Lyu, Z., Zhang, Y., Schneider, B., Lu, Y., Yue, X., and Chen, W. ScholarCopilot: Training large language models for academic writing with accurate citations. *arXiv preprint arXiv:2504.00824*, 2025a. doi: 10.48550/arXiv.2504.00824. URL https://arxiv.org/abs/2504.00824.

Wang, Z., Danek, B., and Sun, J. Biodsa-1k: Benchmarking data science agents for biomedical research, 2025b. URL https://arxiv.org/abs/2505.16100.

Wang, Z., Wang, X., Lee, S., and Xu, X. ARISE: Agentic rubric-guided iterative survey engine for automated scholarly paper generation. *arXiv preprint arXiv:2511.17689*, 2025c. doi: 10.48550/arXiv.2511.17689. URL https://arxiv.org/abs/2511.17689.

Wei, J., Yang, Y., Zhang, X., Chen, Y., Zhuang, X., Gao, Z., Zhou, D., Wang, G., Gao, Z., Cao, J., Qiu, Z., He, X., Zhang, Q., You, C., Zheng, S., Ding, N., Ouyang, W., Dong, N., Cheng, Y., Sun, S., Bai, L., and Zhou, B. From AI for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111*, 2025. doi: 10.48550/arXiv.2508.14111.

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., and Ha, D. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025. doi: 10.48550/arXiv.2504.08066. URL https://arxiv.org/abs/2504.08066.

Yan, X., Feng, S., Yuan, J., Xia, R., Wang, B., Zhang, B., and Bai, L. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629*, 2025. URL https://arxiv.org/abs/2503.04629.

Zhang, W., Li, X., Zhang, Y., Jia, P., Wang, Y., Guo, H., Liu, Y., and Zhao, X. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*, 2025. doi: 10.48550/arXiv.2508.12752.

Zhu, M., Weng, Y., Yang, L., and Zhang, Y. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, 2025. doi: 10.18653/v1/2025.acl-long.1420. URL https://aclanthology.org/2025.acl-long.1420/.

A. Evidence Tiers and Boundary Cases

We use evidence tiers to distinguish a system's visibility from the strength of claims that can be made about its architecture and reliability. The tiers qualify how examples are used in the main text; they are not intended as a ranking of scientific importance. Table 2 summarizes the four tiers used throughout the paper and the kinds of claims each tier supports.

Table 2. Evidence tiers used for representative systems and boundary cases.

Tier	Evidence source	Claims supported	Representative use
T1	Peer-reviewed archival publication or accepted conference/journal paper	Stronger claims about the reported architecture, evaluation protocol, and observed behavior, limited to what the paper documents	Used as core evidence where artifact, verifier, and evaluation are sufficiently described.
T2	Preprint or technical report, especially when accompanied by code, data, benchmarks, or detailed system documentation	Architectural and empirical claims can be discussed, but reliability and generality should be caveated	Used for fast-moving systems where public technical detail is sufficient for framework placement but not final reliability assessment.
T3	Official documentation, repository, product page, or public technical description without full reproducible evaluation	Claims about declared functionality, workflow role, and intended artifact; weak support for reliability or internal architecture	Used mainly for boundary cases and deployment-pattern evidence.
T4	Informal announcement, demo, blog post, or public claim without reproducible technical detail	Visibility and positioning only; not enough for strong claims about system behavior	Used only to explain why a visible system is excluded from the core comparison or treated cautiously.

B. Extended Diagnostic Coding

Table 3 gives a compact coding of representative systems and system families. The table is not an exhaustive catalog. Its purpose is to show how the same framework can be applied across different artifact types, verifier regimes, evidence tiers, and boundary statuses.

C. Benchmark-to-Verifier Mapping

Table 4 summarizes how benchmark families map to artifact types, verifier proxies, and residual limits. The table supports the main paper's claim that there is no single metric of generic AI scientist quality: benchmarks evaluate different artifacts under different verification assumptions.

Table 4. Benchmark families mapped to artifact types, verifier proxies, and remaining limits.

Benchmark family	Examples	Primary artifact	Verifier / proxy	What remains under-tested
Local grounding / claim support	ALCE, CITEBENCH, HALOGEN (Gao et al., 2023; Funkquist et al., 2023; Ravichander et al., 2025)	Citation-supported passage or atomic scientific claim	Citation support, source entailment, or hallucination/support checking	Global document coherence, corpus coverage, and disagreement handling.
Survey synthesis evaluation	SURVEYBENCH (Sun et al., 2025)	Survey or literature review	Reader-oriented or rubric-based assessment of structure, answerability, and informativeness	Whether synthesis provides reliable research guidance beyond local source support.
Execution-backed science tasks	PAPERBENCH, SCIENCEAGENTBENCH, FRONTIERSCIENCE (Starace et al., 2025; Chen et al., 2025b; OpenAI, 2025)	Analysis, reproduction, task solution, or scientific reasoning trace	Code, data, benchmark harness, expert rubric, or external task validator	Whether generated manuscripts state only what execution or expert validation supports.
Review and critique evaluation	MMREVIEW; LLM-review usefulness studies (Gao et al., 2025b; Liang et al., 2024; Thakkar et al., 2025)	Review, feedback, or critique	Human agreement, usefulness, actionability, or review-quality proxy	Whether critique detects consequential scientific error rather than improving perceived quality.

Table 3. Extended diagnostic coding of representative systems and system families. "Core" means central to our comparison set; "boundary" means useful for understanding verification regimes but not always manuscript-first; "adjacent" means relevant to deployment but insufficiently documented for core technical comparison.

System or family	Status	Primary artifact	Verifier / proxy	Tier	Safe interpretation
SCHOLARCOPILOT (Wang et al., 2025a)	Core	Section or passage	Retrieved sources and claim–citation checks	T2	Local drafting co-author; does not establish survey-level coverage or global manuscript validity.
AUTOSURVEY, SURVEYX, SURVEYFORGE (Wang et al., 2024; Liang et al., 2025; Yan et al., 2025)	Core	Survey or literature review	Corpus construction, source support, and synthesis checks	T2	Synthesis co-author; useful for literature organization, but vulnerable to coverage gaps and smoothed disagreement.
DATA-TO-PAPER (Ifargan et al., 2025)	Core	Full paper or report	Code, data, numerical traceability, and provenance	T1	Strong example of traceable manuscript claims; breadth of autonomous exploration is not the main claim.
THE AI SCIENTIST, AGENT LABORATORY (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025)	Core / boundary by version	Full paper or research pipeline output	Execution traces, logs, benchmarks, and reviewer-like checks	T2	More autonomous research-loop behavior; claims must be interpreted relative to execution depth and validation strength.
AGENTREVIEW, SWIF^{2}T, DEEPREVIEW (Jin et al., 2024; Chamoun et al., 2024; Zhu et al., 2025)	Core	Review, feedback, or rebuttal support	Human agreement, usefulness, actionability, and review-quality proxies	T1–T2	Reviewer assistant; critique quality should not be mistaken for scientific verification.
Formal proof agents (Feng et al., 2026; Ospanov et al., 2025)	Domain evidence	Formal proof or formalized claim	Proof-assistant feedback after formalization	T2	High autonomy is more plausible after formalization; informal prose proofs remain weak without machine checking.
COSCIENTIST, A-LAB, GRACE (Boiko et al., 2023; Szymanski et al., 2023; Hill & Ryoo, 2026)	Boundary / regime evidence	Protocol, experiment, simulation, or lab workflow	Instrument, simulator, or characterization loop	T1–T2	Shows tool-mediated verification pressure; local tool success is not the same as final physical truth.
AI CO-SCIENTIST, BIODISCOVERYAGENT (Gottweis et al., 2025; Roohani et al., 2025)	Boundary / regime evidence	Hypothesis, proposal, or ranked experiment	Literature, existing data, and later empirical validation	T2	Founder-like triage before validation; should not be described as completed biological discovery unless externally validated.