Gresham's Law for Conference Submissions: Adopting a Tiered Contribution Taxonomy for the Agentic Research Era

Raeid Saqur[1,2] **Tom Klose**[1]
[1]Mathematical Institute, University of Oxford
[2]Vector Institute
{raeid.saqur, tom.klose}@maths.ox.ac.uk

Abstract

This position paper argues that the conference review system is structurally mismatched to agentic research and that the mismatch is already producing adverse selection. The system was designed around two assumptions: that a static paper is the atomic unit of contribution, and that the marginal cost of producing one such unit is high enough that a single uniform review track can triage quality. Agentic research systems violate both. Building on Akerlof's *market-for-lemons* argument, we argue that under finite reviewer capacity and growing indistinguishability between human- and agent-produced submissions, the equilibrium share of agent-produced contributions grows without bound, eventually crowding out human-driven submissions. We propose three coupled reforms: (i) contribution-class declarations distinguishing agent-as-tool, agent-as-co-author, and agent-as-author; (ii) verifiable process artifacts (e.g. trajectory logs, versioned pipelines, and experiment DAGs) as a strictly stronger reproducibility primitive than code release; and (iii) a distinct agent-assisted research track with review criteria calibrated to process-centric contributions. The reforms are incremental, falsifiable, and compatible with existing infrastructure. They make the resource asymmetry between agent-running and non-agent-running submitters explicit and auditable rather than hidden behind a uniform, homogenous track.

1 Introduction

In March 2025, a fully autonomous machine learning system produced a paper that received an average reviewer score of 6.33 at the ICLR 2025 ICBINB workshop, exceeding the workshop's average acceptance threshold; the authors withdrew the manuscript before publication only because the field had not yet decided whether to publish it [Yamada et al., 2025, Sakana AI, 2025]. In the same year, an alignment research agent at a frontier lab closed 0.97 of the human–vs.–strong-model performance gap on a real weak-to-strong supervision problem in five days, while a comparable team of human researchers reached 0.23 in seven [Wen et al., 2026]. By early 2026, an open-source single-GPU loop fitting on a laptop autonomously modifies, trains, and selects ML training code overnight without further human input [Karpathy, 2026]. None of these systems existed two years ago.

Position. **The conference review system was designed for human-authored static artifacts which enhance human understanding, and the assumptions that make it work are now violated. NeurIPS should adopt a tiered contribution taxonomy with verifiable process artifacts and a distinct agent-assisted track, rather than continuing to triage all submissions through a single uniform pipeline.**

Preprint.

The case has three complementary pieces. The first is structural: papers are snapshots; agentic research is a trajectory. Standard reproducibility, attribution, and benchmark norms were calibrated for the snapshot, not the trajectory, and they fail in predictable ways when the trajectory becomes the operative artifact. The second is economic: the marginal cost of producing an artifact resembling a paper is approaching zero for a growing class of contribution types, while reviewer capacity remains bounded. Under bounded distinguishability between human- and agent-produced artifacts, this is a textbook 'adverse-selection' setting. Akerlof's market for lemons [Akerlof, 1970] predicts what follows: the lower-cost type crowds out the higher-cost type. The single-track conference is the lemons market. The third is epistemological: Papers have never been an end but only a means to further the understanding of their human readers. To stay aligned with that goal, the fast-improving capacities of agentic systems require a change in evaluation metric focused on the deliberation process itself.

Contributions.

- A semi-formal model (Section 4) of the adverse-selection dynamic in conference review under falling paper-production costs and de-facto indistinguishability of human- and agent-driven articles for reviewers.
- A four-axis structural diagnosis (Section 3) of how static-paper review norms fail under agentic research, distinct from — and not reducible to — generic concerns about LLM-assisted writing.
- A concrete, incremental reform proposal (Section 5): contribution-class declarations, mandatory process artifacts for agent-co-authored and agent-authored work, and a separately reviewed agent-assisted track. The proposal is falsifiable: each component can be piloted without committing to the full reform.
- An adversarial defense (Section 7) against five steelmanned objections, including the strongest: that the proposal entrenches incumbents.

This is not a survey of AI-assisted science, and it is not a denunciation of agentic research. It is an argument that a community whose review infrastructure assumes a particular cost structure should reform the infrastructure when that cost structure changes.

2 Background: Agentic Research in 2026

We define an agentic research system as one in which an autonomous controller — typically an LLM with access to code execution, retrieval, and a memory — selects intermediate research actions (hypothesis generation, experiment design, code modification, analysis, write-up) without per-step human approval. Three regimes are operationally distinct.

Agent-as-tool. A human researcher uses an LLM for writing assistance, code completion, search, or proof checking. The human selects every intermediate step. This is the regime addressed by existing LLM-use disclosure policies [ICML, 2026, NeurIPS, 2026].

Agent-as-co-author. An agent autonomously executes one or more substantive research sub-tasks — e.g., proposes and runs an experiment grid, or drafts a method given a hypothesis — under post-hoc human oversight. Recent agent laboratories [Schmidgall et al., 2025] and idea-generation studies [Si et al., 2024] occupy this regime.

Agent-as-author. An agent drives the full loop: hypothesis to experiment to manuscript. The AI Scientist [Lu et al., 2024, Yamada et al., 2025], autoresearch [Karpathy, 2026], and Anthropic's automated weak-to-strong researcher [Wen et al., 2026] are proof that such systems do already exist right now.

Agent-tractable contribution. A contribution type whose production pipeline an agent can complete end-to-end at quality competitive with the lower decile of accepted human submissions, given current capability frontiers. Agent-tractable contributions today include: incremental algorithmic variants on standard benchmarks, ablation-style studies, and template-following empirical follow-ups [Yamada et al., 2025, Si et al., 2024]. Agent-intractable contributions today include: novel formal frameworks, contributions requiring physical apparatus or human-subject data, and theoretical proofs requiring nontrivial original construction. The taxonomy is not a fixed boundary; it is calibrated to the current frontier and expected to be versioned (Section 7).

3 The Structural Failure of Static-Paper Norms

The paper-as-unit assumption breaks along four axes when confronted with agentic research. The breakages are not addressable by minor reviewer-instruction updates; they are mismatches between the artifact reviewers see and the artifact reviewers should be evaluating.

(i) Static vs. dynamic. A paper is a snapshot of a frozen pipeline. An agentic system is a trajectory through a search space: branching hypotheses, discarded experiments, mid-run prompt revisions, retrieval calls, intermediate model checkpoints. The write-up condenses one path through this space; reviewers cannot evaluate the rest. The relevant question for an agent-authored submission is not "did the reported result hold?" but "how was this path selected from the space the agent searched?" — and that question is unanswerable from the static artifact.

(ii) The reproducibility illusion. The NeurIPS reproducibility checklist [Pineau et al., 2021] is calibrated to a model in which code, data, and hyperparameters reconstitute the result. For an agent-authored submission, this is necessary but not sufficient: the agent's tool calls, prompts, intermediate reasoning, retry decisions, and chosen-vs.-discarded branches are the operative reproducible unit. A run replayed from frozen code does not exercise the agent's selection procedure — it exercises only the procedure's terminal output.

(iii) Attribution breakdown. Existing authorship norms assume identifiable human contributors who can take responsibility for each claim [Stokel-Walker and Van Noorden, 2023, ICML, 2026]. When an agent generates the hypothesis, designs the experiment, runs the analysis, and drafts the manuscript, the human listed as author either (a) has accountability without contribution, or (b) the contribution is the prompt and oversight protocol, which existing review formats do not solicit. Both ICML 2026 and NeurIPS 2026 explicitly disallow listing LLMs as authors [ICML, 2026, NeurIPS, 2026]; this resolves the credit question by fiat but not the substantive contribution question.

(iv) Benchmark overfitting masked by static review. An agent optimizing an open-loop reward signal — e.g., validation loss on a fixed benchmark — will exploit the benchmark in ways that do not generalize [Recht et al., 2019, Sculley et al., 2018]. Static-paper review evaluates the reported result against the benchmark; it cannot evaluate the search procedure that produced it. Process-centric review can: a trajectory log reveals whether the agent searched orthogonal hypotheses or merely incremented a single hyperparameter axis until the benchmark moved.

These four failures are coupled. The reproducibility illusion enables benchmark overfitting; attribution breakdown shields it; static-vs.-dynamic prevents reviewers from seeing it. The fix to any one axis is partial unless the artifact under review changes.

4 The Gresham's Law Argument

We now address the adverse-selection claim. The argument has the shape of Akerlof's market for lemons [Akerlof, 1970]: when buyers (reviewers) cannot reliably distinguish high-quality from low-quality goods[1], and the cost of producing a low-quality good is much lower than the cost of producing a high-quality good, the market equilibrates toward low-quality goods crowding out high-quality ones — a Gresham's-Law dynamic in disguise [Mundell, 1998].

Submission types. For a fixed conference cycle, we may sort the submissions into two categories: either they are human-driven (H) or agent-tractable (A, cf. Section 2). Producing a single submission incurs a marginal cost per unit which, we assume, is strictly smaller for agent-driven than that for human-driven works; given the recent advances of AI systems, it is plausible to assume that the cost for the former is monotonically decreasing over time.

Reviewers. In contrast, the number of reviewers as well as their reviewing capacity remains approximately constant. According to current detection literature [Sadasivan et al., 2023, Liang et al., 2024, Yu et al., 2025, Rao et al., 2025], they cannot reliably distinguish the type (H or A) of each submission they receive.

Metaproposition 1 (Gresham's Law for Conference Submissions)**.** *Suppose the marginal cost to produce an agent-tractable paper monotonically decreases to zero as agentic capabilities increase*

[1] In case of Neurips, the reason emanates from capacity vs. abilities of reviewers.

while the marginal cost for human-driven contributions is bounded below. Suppose further that the average quality of type-A submissions is at or above the marginal acceptance threshold of the conference (see paragraph at the end of this section for an empirical anchor).

If both submission types are treated and reviewed equally, the following dynamics will be observed over time:

1. *The share of submissions of type-A will equilibrate to one over time, slowly crowding out type-H submissions.*
2. *Capable human researchers reallocate their effort to venues which have a more beneficial acceptance probability that rewards their cost-to-quality ratio.*

What the proposition does and does not say. It does not say agent-produced submissions are bad. It says that under bounded distinguishability and falling production cost, the share dynamics are independent of intrinsic quality. The system equilibrates to a state where the median submission costs less and high-cost researchers self-select out — regardless of whether the agent-produced contributions are scientifically valuable. This is precisely Akerlof's failure mode: the market does not crash because the goods are bad; it crashes because asymmetric information prevents prices (here, acceptance probabilities) from rewarding the costlier type.

The empirical indistinguishability assumption. The argument is valid if and only if human reviewers can, indeed, not reliably distinguish submission types. Three lines of evidence support this. First, theoretical results on AI-text detection establish that reliable detection has a fundamental upper bound determined by the total-variation distance between human and AI text distributions, and that this bound is approached by paraphrase attacks [Sadasivan et al., 2023]. Second, large-scale audits of AI conference reviews in 2023–2024 estimate non-trivial fractions of LLM-modified review text that escaped detection at submission time [Liang et al., 2024]. Third, recent benchmarks for detecting LLM-generated peer reviews find that off-the-shelf detectors fail under standard adversarial paraphrasing, and that watermark-based methods rely on producer cooperation that adversaries simply opt out of [Yu et al., 2025, Rao et al., 2025]. The asymmetry is therefore favoring producers: Detection improvements help, but they do not close the gap.

Empirical anchor. The AI Scientist v2 ICBINB submission [Yamada et al., 2025, Sakana AI, 2025] shows that the proposition's premise is valid: an agent-produced submission cleared the workshop's marginal acceptance threshold. The submission was withdrawn by the producers, not rejected by the reviewer, and this decision was predicated on community norms that do not yet exist as policy. This is the precise tension the reform proposal addresses.

5 Proposed Reform: Tiered Taxonomy and Process Artifacts

The structural and adverse-selection diagnoses point to the same intervention: change the artifact under review and the criteria applied to it. We propose three coupled mechanisms. Each is independently implementable; the full effect requires all three.

5.1 Contribution-class declaration

At submission, authors declare (i) which agent regime applies (tool, co-author, author; see Section 2) and (ii) which contribution type the work claims (algorithmic, empirical, theoretical, dataset, framework, position, etc.). The declaration is a structured field, not a free-text disclosure. Existing LLM-use checkboxes [ICML, 2026, NeurIPS, 2026] do not separate "used Copilot for one method's draft" from "the method, experiments, and write-up were produced by an autonomous loop"; the former is an editing tool, the latter is a different artifact and should be reviewed differently.

The declaration alone is not gaming-proof: a self-reported field gives the gaming author an obvious lever. Self-declaration is therefore necessary but not sufficient, and must be coupled with mechanism 5.2, which is qualitatively harder to fake.

5.2 Verifiable process artifacts

For agent-co-author and agent-author regimes, submissions must include:

- A **trajectory log**: agent tool calls, prompts, intermediate reasoning traces (where exposed by the underlying system), and retry/branching decisions, in a structured format.
- A **versioned pipeline**: a containerized or specification-level reproduction of the production environment including model versions, retrieval indices, and any non-frozen randomness.
- An **experiment DAG**: the directed graph of hypotheses generated, experiments run (including discarded ones), and decisions made between them.

This is strictly more information than the current code-and-data standard [Pineau et al., 2021]: the trajectory exposes the search procedure, not just the search winner. A reviewer can replay the agent's selection over the full DAG, not just the terminal artifact, and can ask whether the agent searched orthogonal hypotheses or only one axis (the failure mode in Section 3, axis (iv)).

The artifact requirement is qualitatively harder to game than a checkbox: forging a trajectory log requires producing a self-consistent search history with the right statistical signature of an agent run, and conferences can spot-check by re-executing pipeline stages. We do not claim full unforgeability — only that the cost asymmetry shifts back toward the producer.

5.3 Agent-assisted research track

We propose a distinct submission track for agent-co-author and agent-author submissions. Reviewers for this track are briefed on agentic system capabilities and on the artifact format. Review criteria are adapted on three points: (a) novelty is evaluated at the level of the search procedure (what hypotheses did the agent consider, and why is that space novel?), not only at the level of the search winner; (b) reproducibility is evaluated against the trajectory replay, not only against the code; (c) generalization is evaluated by asking whether the agent's selection rule, applied to a held-out hypothesis space, produces meaningfully different outputs — a process-level analogue of held-out evaluation. The track has its own acceptance criteria and its own decision norms, separately from the main track.

Implementation questions — who verifies artifacts, how to handle partially-released model APIs, how to evaluate when the agent's reasoning is not exposed — are real but bounded. The reproducibility checklist faced exactly the same questions in 2019–2020 [Pineau et al., 2021]; they were answered iteratively.

5.4 Process as contribution

Figure 1 contrasts the two models. Under the static model, the submission is a terminal artifact. Under the process-centric model, the submission is the trajectory plus the terminal artifact, with the trajectory itself a first-class object of review. This is not an extension of the existing reproducibility framework; it is a different unit of contribution. A novel research DAG — a search procedure that explores a hypothesis space no human group had organized before — is a contribution per se, evaluable by replayability rather than by the existing notion of reproducibility.

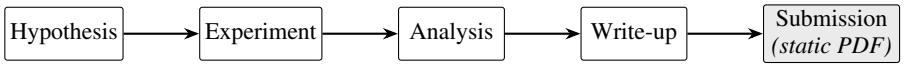
6 Implications

Peer review. Reviewers for the agent-assisted track require briefing on agentic system capabilities and on artifact-replay tooling; this is a non-negligible training cost but tractable on the timescale of a single conference cycle, comparable to the rollout of the reproducibility checklist [Pineau et al., 2021]. Alternatively, the track could be further automated by delegating the review to agents as well – something that is already an optional feature for NeuRIPS main track.

Open science. Process artifacts extend the open-science norm beyond code and data. A research DAG and a trajectory log are public-good infrastructure: they let third parties audit, reproduce, and extend the search procedure. Standard code release does not provide this.

Industry vs. academia. A common objection is that agent-track artifact requirements favor compute-rich actors. The current uniform model already favors them: an industry lab running an autonomous agent overnight produces submissions at a cost their academic counterparts cannot match. The reform

(a) Static-paper model. Reviewer sees only the terminal artifact.



(b) Process-centric model. Submission bundles the trajectory, not just the terminal write-up.

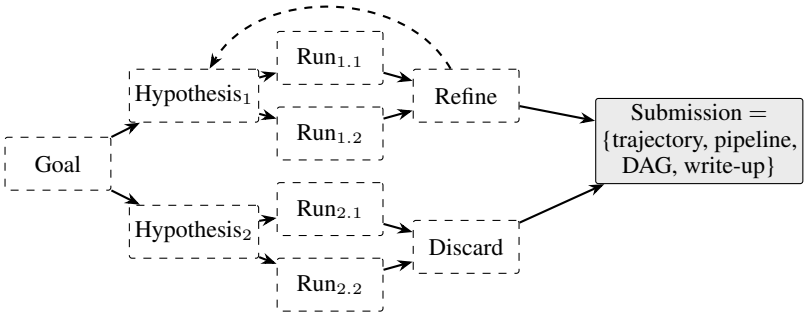


Figure 1: Two models of scientific contribution. **(a)** The static-paper model treats a linear pipeline's terminal artifact as the unit of review. **(b)** The process-centric model treats the agent's branching trajectory — including discarded paths — as the auditable artifact, with the write-up as one of several first-class outputs. Process-centric review evaluates the search procedure, not only the search winner.

makes this asymmetry explicit and auditable rather than hidden. A small academic group submitting to the main human-author track competes on terms calibrated to its cost structure; a frontier-lab submission to the agent-author track is reviewed against criteria that recognize its cost structure. Both are legible.

Incentive realignment. A contribution-class declaration changes what is profitable to submit. A researcher with a genuinely novel process design has a clear submission home; a researcher with a five-minute prompt-and-paper-shaped artifact does not benefit from miscategorizing it as agent-assisted-author. The mechanism rewards process novelty and penalizes nothing that the current system does not already implicitly penalize.

7 Alternative Views and Objections

Objection 1: This misses the point. Papers have never been the end goal: Advancing human knowledge has and standard benchmarks are a proxy for that. If agent-driven papers improve on those benchmarks, there is no reason to distinguish them from human-driven contributions. Our objection: Benchmarking is fragile. Until recently, it is usually a new idea that leads to an improvement on a chosen benchmark; it is unclear if that implicit requirement still holds. Even in the field of Mathematics, where one obvious benchmark is logical correctness, there is an ongoing lively debate about the interpretations of recent agent-driven advances [Kontorovich, 2025, Avigad, 2026, Bessis, 2026]. In the absence of a definitive answer, it seems prudent to distinguish human- and agent-driven submissions with a focus on process, not least to enhance our understanding of agentic research contributions as an object in themselves.

Objection 2: This is premature. "Agentic systems are not yet capable enough to constitute a real threat to conference integrity." The most defensible version of this view is that the AI Scientist v2 ICBINB submission [Yamada et al., 2025] is a workshop-track outlier, not a top-tier-conference submission, and that current detectors [Liang et al., 2024] keep AI/human-paper distinguishability high enough in practice. Our response: the question is whether the infrastructure should be built now or in crisis mode later. The reproducibility checklist took two years to design; a tiered taxonomy will take comparable time. If the threat is two years away and the response time is two years, the build window is now. The ICBINB event already produced an ad-hoc community decision (the producers withdrew); ad-hoc precedents calcify into bad policy.

Objection 3: The agent-tractable / agent-intractable distinction will collapse as agents improve. If sufficiently capable agents eventually match human researchers across all dimensions, the tax-

onomy axis is unstable. Our response: the taxonomy is calibrated to the current capability frontier and is explicitly versioned. Other regulatory frameworks (FDA trial phases, financial-disclosure requirements) are versioned to changing technology and have not been undone by their evolution. A versioned taxonomy is a feature: it forces the community to re-examine the boundary periodically rather than freezing a snapshot.

Objection 4: The taxonomy will be gamed (i.e. susceptibility to Goodhart's Law) Researchers will misclassify their work to avoid scrutiny or land in a more favorable track. Our response: self-declaration is not the enforcement mechanism; verifiable process artifacts are. Gaming a structured trajectory log is qualitatively harder than gaming a checkbox, because the artifact has a statistical structure (branching factor, retry frequency, reasoning-step length distribution) that is detectable by spot-check. The submission cost of forging consistent artifacts approaches the submission cost of running the agent. We do not claim full unforgeability; we claim that the cost asymmetry is moved in the right direction.

Objection 5: This entrenches incumbents. A separate track with stricter artifact requirements favors compute-rich labs. Smaller groups are penalized. This is the strongest objection. Our response has two parts. First, the current uniform model already entrenches incumbents: a frontier lab that produces ten agent-authored submissions in the time it takes a small group to produce one is already winning the volume game; the uniform track simply hides this. Second, separating tracks creates a venue where small groups compete only against other small groups (the human-author main track) and where their cost structure is rewarded; without separation, they compete in a single pool against agent-produced submissions whose cost they cannot match. The reform improves, not worsens, the small-group position.

8 Conclusion

The conference review system was designed for a cost regime that no longer holds. Static-paper review evaluates a snapshot when the operative artifact is a trajectory; uniform review triages all contribution types in a queue whose composition is shifting when reviewers cannot reliably distinguish agentic and human-driven submissions. Both pieces motivate the same reform: a tiered contribution taxonomy with verifiable process artifacts and an agent-assisted track. The reform is incremental, falsifiable, and compatible with existing infrastructure. The smallest concrete near-term step — one that does not require committing to the full reform — is for NeurIPS to add a structured contribution-class declaration as a submission field for the next cycle, and to begin piloting the artifact format on a workshop or special-issue scale. The cost is small; the option value of the resulting data is large. The cost of waiting is that the next ad-hoc precedent, like the ICBINB withdrawal, will be made by individual producers rather than by community policy.

References

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. The Quarterly Journal of Economics, 84(3):488–500, 1970.

Jeremy Avigad. Mathematicians in the Age of AI. arXiv preprint arXiv:2603.03684, 2026.

David Bessis. The fall of the theorem economy. Substack, 2026. URL `https://davidbessis.substack.com/p/the-fall-of-the-theorem-economy`.

ICML. ICML 2026 publication ethics and LLM policy. Conference website, 2026. URL `https://icml.cc/Conferences/2026/PublicationEthics`.

Andrej Karpathy. autoresearch: nanochat-based agentic ML research loop. GitHub repository, 2026. URL `https://github.com/karpathy/autoresearch`. Accessed May 2026.

Alex Kontorovich. The Shape of Math To Come. arXiv preprint arXiv:2510.15924, 2025.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, and James Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. arXiv preprint arXiv:2403.07183, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292, 2024.

Robert Mundell. Uses and abuses of Gresham's law in the history of money, 1998.

NeurIPS. Call for position papers, NeurIPS 2026. Conference website, 2026. URL https://neurips.cc/Conferences/2026/CallForPositionPapers.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). Journal of Machine Learning Research, 22(164):1–20, 2021.

Vishisht Rao, Vinu Sankar Sadasivan, and Soheil Feizi. Detecting LLM-generated peer reviews. arXiv preprint arXiv:2503.15772, 2025.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In International Conference on Machine Learning (ICML), 2019.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156, 2023.

Sakana AI. The AI scientist generates its first peer-reviewed scientific publication. Sakana AI blog, March 2025. URL https://sakana.ai/ai-scientist-first-publication/. Submission accepted to ICLR 2025 ICBINB Workshop.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent Laboratory: Using LLM agents as research assistants. arXiv preprint arXiv:2501.04227, 2025.

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. In ICLR Workshop Track, 2018.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. arXiv preprint arXiv:2409.04109, 2024.

Chris Stokel-Walker and Richard Van Noorden. What ChatGPT and generative AI mean for science. Nature, 614:214–216, 2023.

Jiaxin Wen, Liang Qiu, Joe Benton, Jan Hendrik Kirchner, and Jan Leike. Automating weak-to-strong generalization research. Anthropic Alignment Science blog, 2026. URL https://alignment.anthropic.com/2026/automated-w2s-researcher/.

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

Sungduk Yu et al. Is your paper being reviewed by an LLM? benchmarking AI text detection in peer review. arXiv preprint arXiv:2502.19614, 2025.