SCIREVIEW: A Benchmark for Evaluating Frontier AI for Scientific Review

Anonymous Authors[1]

Abstract

Large language models are increasingly positioned as AI scientists, yet existing evaluations focus on hypothesis generation, coding, experimentation, or reproducibility rather than on a capability that is central to scientific reliability: reviewer-style error detection, critique and feedback. We introduce SCIREVIEW, a benchmark that evaluates whether frontier models can read a realistic research writeup and identify locally plausible, high-consequence conceptual errors. Each task begins from an expert-authored research text; a domain expert then injects a small set of natural errors and provides gold rationales, while the protocol explicitly excludes trivial lookup mistakes, unsupported falsehoods, and items that break internal coherence. Errors are calibrated into *baseline* and *challenging* difficulty tiers via adversarial filtering against multiple frontier models. We evaluate frontier models GPT-5.4, Gemini 3.1 Pro, and Claude Opus 4.6 under five complementary scoring regimes and three qualitative axes (helpfulness, correctness, alignment). No model achieves perfect error recovery on more than a single item; the strongest uncorrected recaller (GPT-5.4, 62% average recall) collapses to 10% once a single false positive disqualifies the run. SCIREVIEW complements recent AI-for-science benchmarks by measuring the capability that matters most for preventing researchers from acting on faulty premises before publication.

1. Introduction

Large language models (LLMs) are increasingly positioned as research assistants that can formulate hypotheses, write code, design experiments, summarize literature, and draft manuscripts (Luo et al., 2025; Zhou et al., 2025; Lu et al., 2024; Gottweis et al., 2025). This trend has fueled the broader idea of an "AI scientist": a system that supports, or even partially automates, scientific work (Lu et al., 2024; Gottweis et al., 2025). Yet scientific usefulness is not only about generating ideas or executing workflows. It is also about *catching consequential mistakes* before they propagate.

That capability matters because the reliability of the scientific record is already under pressure. In one widely cited survey, more than 70% of researchers reported having failed to reproduce another scientist's results (Baker, 2016); large-scale replication efforts have found effect sizes in preclinical biology to be systematically overstated (Begley and Ellis, 2012; Errington et al., 2021); and analyses of the published literature suggest that the base rate of false findings may be far higher than commonly assumed (Ioannidis, 2005). A recent consensus report from the National Academies underscores that these concerns span virtually every empirical discipline (National Academies, 2019). At the same time, frontier LLMs are being incorporated into exactly the workflows—experimental planning, grant writing, and peer review—where undetected errors are most costly (Bauchner and Rivara, 2024; Liang et al., 2024b). If AI systems are to participate meaningfully in research, they must do more than produce fluent scientific text: they must reliably identify when a method overreaches, when an assumption silently fails, when an interpretation is invalid, or when a local technical claim breaks the larger argument.

Recent benchmarks have substantially improved evaluation for research-adjacent tasks—machine learning experimentation (Huang et al., 2024; Chan et al., 2024), data-driven discovery (Majumder et al., 2024), scientific coding (Tian et al., 2024), workflow-level agents (Chen et al., 2024), computational reproducibility (Siegel et al., 2024), paper replication (Starace et al., 2025), citation attribution (Press et al., 2024), scientific claim verification (Wadden et al., 2020), and manuscript verification from published errors (Son et al., 2025). These are all valuable, but they leave a specific gap: *reviewer-style detection of embedded errors in realistic, in-progress scientific writeups*.

We introduce SCIREVIEW to fill this gap. The paper makes

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop at the 43^{rd} International Conference on Machine Learning (ICML 2026).

three contributions. First, we formulate reviewer-style scientific error detection as a distinct benchmark task for AI scientists and describe a construction protocol that emphasizes realistic expert writing, natural error injection, adversarial difficulty calibration, and explicit exclusion of trivial artifacts. Second, we define a multi-regime scoring framework that reports recall under three aggregation schemes crossed with two false-positive treatments, alongside expert-rated qualitative axes, allowing readers to examine the accuracy–over-flagging trade-off that is central to this task. Third, we report results showing that frontier models achieve only modest gold-error recall and near-zero exact-set match—and that model rankings invert once false positives are penalized, revealing a recall–precision trade-off with direct operational implications.

2. Related Work

AI scientists and research-support systems. Recent systems increasingly frame LLMs as components of broader research workflows. Lu et al. (2024) introduce an end-to-end framework for automated ML research, while Gottweis et al. (2025) present a multi-agent system for hypothesis generation in biomedicine. Surveys such as Luo et al. (2025) and Zhou et al. (2025) organize this space across hypothesis generation, experimentation, scientific writing, and peer review. SCIREVIEW focuses on one narrow but consequential capability within this larger landscape: scientific error detection during review.

Benchmarks for scientific coding, experimentation, and reproducibility. MLAgentBench studies machine learning experimentation (Huang et al., 2024). DiscoveryBench formalizes multi-step data-driven discovery (Majumder et al., 2024). SciCode measures scientific problem solving through research coding tasks (Tian et al., 2024). ScienceAgentBench evaluates agents on data-driven discovery workflows (Chen et al., 2024). CORE-Bench evaluates computational reproducibility (Siegel et al., 2024), MLE-bench focuses on ML engineering through Kaggle competitions (Chan et al., 2024), and PaperBench measures replication of recent AI papers from scratch (Starace et al., 2025). These benchmarks emphasize execution, implementation, or reproduction rather than error detection in continuous scientific prose.

Benchmarks for scientific verification and scholarship. SciFact evaluates whether systems can retrieve evidence that supports or refutes scientific claims (Wadden et al., 2020; 2022). CiteME studies citation attribution from claim-bearing excerpts (Press et al., 2024). These tasks are narrower than reviewer-style manuscript critique: they evaluate atomic claims or citations rather than the validity of an extended research argument. FLAWS (Xi et al., 2025) scales error detection to 713 paper–error pairs by using LLMs to *insert* claim-invalidating errors into peer-reviewed papers; the best model achieves 39.1% identification accuracy at k=10. PaperAudit-Bench (Tu et al., 2026) applies a similar LLM-injection strategy to machine-learning papers. SPOT (Son et al., 2025) is the closest benchmark: it evaluates automated verification using 83 manuscripts with real, high-severity errors from errata or retractions, reporting that no frontier model exceeds 21.1% recall or 6.1% precision. SCIREVIEW differs from SPOT and the LLM-injection benchmarks on three axes: (i) errors are *written by domain experts*, not synthesized by an LLM or harvested from retractions; (ii) the task operates on *research proposals and methodology passages* rather than already-published papers; and (iii) items are *adversarially filtered* so that each retained task contains errors that all tested frontier models missed at construction time.

Open-ended expert benchmarks and scientific judgment. Frontier-difficulty QA benchmarks including GPQA (Rein et al., 2024), MATH (Hendrycks et al., 2021), FrontierMath (Glazer et al., 2024), HLE (Center for AI Safety et al., 2026), and ATLAS (Liu, 2025) measure a model's *knowledge* and *problem-solving* on research-level content. They are complementary to SCIREVIEW, which evaluates a model's ability to critically read already-written research prose. Step-level verification benchmarks such as Hard2Verify (Pandit et al., 2025) and ProcessBench (Zheng et al., 2024) examine verifier reliability on machine-generated mathematical proofs, whereas SCIREVIEW examines it on naturalistic human-written prose. Proof of Time (Ye et al., 2026) evaluates how models judge scientific ideas by forecasting downstream outcomes; it studies whether models can assess which ideas *may matter*, whereas SCIREVIEW studies whether a model can detect which claims in a concrete draft are *wrong*.

3. Benchmark Design

3.1. Task Definition

Each SCIREVIEW item is a tuple (T, E, C) where T is a research-grade text passage ($\sim$150–800 words), $E = \{e_1, \ldots, e_k\}$ is a set of errors intentionally injected into T by a domain expert, and $C = \{c_1, \ldots, c_k\}$ is the corresponding set of expert-authored descriptions and corrections. A model reads T and outputs a free-form review that identifies purported problems and, optionally, proposes corrections. The benchmark then maps the review to the gold error set E.

This formulation intentionally mirrors a realistic use case: a scientist asks an AI system to critique a draft proposal, method section, or technical exposition. The central question is whether a frontier model can identify scientifically

Table 1. Positioning SCIREVIEW relative to closely related evaluations. The gap we target is reviewer-style detection of embedded scientific errors in realistic, expert-authored writeups.

Benchmark	Primary object	Main capability	Typical setting	Main distinction from SCIREVIEW
GPQA / HLE (Rein et al., 2024; Center for AI Safety et al., 2026)	Expert questions	Closed-ended expert QA	Verifiable answers	Measures knowledge rather than critique of scientific prose
SciFact / CiteME (Wadden et al., 2020; Press et al., 2024)	Claims or citations	Claim verification / attribution	Retrieval or attribution	Targets atomic scholarly units rather than extended realistic draft arguments
Workflow benchmarks (Tian et al., 2024; Chen et al., 2024; Siegel et al., 2024; Starace et al., 2025)	Code, workflows, papers	Execution, reproduction, replication	Tool use or implementation	Emphasize doing scientific work rather than reviewing it
SPOT (Son et al., 2025)	Published papers with real failures	Manuscript verification	Errata / retractions	Post-publication failures rather than expert-injected draft errors
FLAWS (Xi et al., 2025)	Scientific papers	Error identification and localization	LLM-inserted errors	Automated insertion rather than expert-authored errors
SCIREVIEW (ours)	Expert-authored writeups	Comprehensive reviewer-style error detection and feedback	Draft-time critique	-

meaningful failure points embedded in otherwise plausible technical prose.

The benchmark spans four super-domains, each populated by experts with graduate training: **Science** (biology, chemistry, neuroscience, physics), **Engineering** (aerospace, biomedical, civil, chemical, electrical, mechanical, industrial), **Technology** (data science, IT), and **Mathematics** (pure, applied, statistics). Because expert standards for rigor differ across fields, tasks are authored and graded within a single discipline, and mixed-discipline tasks are excluded.

3.2. Authoring Protocol and Exclusion Criteria

The benchmark design is guided by five principles:

P1: Start from realistic expert writing. Seed texts are authored or selected by domain experts and resemble actual research artifacts—proposals, method sections, technical notes—rather than synthetic puzzles. The aim is to make model failures operationally meaningful for actual scientific workflows.

P2: Inject natural errors, not adversarial trivia. Experts insert a small number of errors that are locally plausible, scientifically consequential, and natural enough that a working researcher could miss them. For every injected error, the expert provides: a precise location, a description of the conceptual mistake, the downstream reasoning affected, and one or more acceptable corrections.

P3: Exclude trivial or poorly grounded items. The authoring protocol distinguishes the desired error class from several weaker alternatives. We explicitly require *conceptual* errors—misunderstandings that propagate through a chain of reasoning—and exclude three patterns that produced unreliable test items during development: bare lookup errors with no downstream role, altered definitions where the downstream logic remains internally coherent, and unsupported false assertions with no local justification. Table 2 summarizes the inclusion and exclusion criteria. Errors that make the passage internally incoherent are also rejected, since they are trivial to detect.

P4: Shape difficulty during authoring. Each candidate task is evaluated once against three frontier models prior to acceptance. A *challenging error* is one missed by all three models; each accepted task must contain at least three. A *baseline error* is one caught by all three; each task is capped at three to retain diagnostic value without trivial saturation. This adversarial filter gives items a mixture of easy and hard signals while ensuring headroom against frontier systems.

Table 2. Included and excluded error patterns in the authoring protocol.

Treatment	Error pattern
Include	Faulty assumption that breaks downstream reasoning
Include	Invalid extrapolation across levels of analysis
Include	Limiting-case or mechanism error that preserves local plausibility
Exclude	Bare lookup fact with no downstream role
Exclude	Altered definition with internally coherent downstream logic
Exclude	Unsupported false claim with no local justification

P5: Keep explanations attached to gold labels. Each item includes an expert list of the injected errors together with rationales that explain why each error matters. These rationales are essential both for evaluation and for downstream analysis of model failures.

3.3. Quality Control

Each candidate task undergoes a review and revision process by an independent expert who rates the task on an ordinal scale; only items rated "Good" enter the evaluation set. Common failure modes rejected during review were: (i) errors that were technically incorrect but did not propagate (lookup errors), (ii) errors indistinguishable from authorial imprecision, and (iii) errors that made the passage internally incoherent (thus trivial to spot). The authoring instructions were iteratively refined to reduce these failure modes.

4. Evaluation Methodology

4.1. Models and Protocol

We evaluate three frontier models: **GPT-5.4** (OpenAI), **Gemini 3.1 Pro** (Google DeepMind), and **Claude Opus 4.6** (Anthropic). Each model receives a uniform prompt instructing it to identify and explain errors, and where possible, propose corrections. Each passage is evaluated 4 times per model (independent runs with default sampling temperature).

4.2. Matching and Scoring

Because model responses are free-form critiques, we employ an LLM-based matcher (following the LLM-as-judge paradigm (Zheng et al., 2023)) to determine which expert-annotated errors each response identified and how many additional non-errors were flagged. We validated matcher reliability on a stratified subsample against domain-expert re-verification; future releases will add systematic human auditing (see Section 7).

Let $M(r_i)$ denote the gold errors matched by response r_i on item i, and $\mathrm{FP}(r_i)$ the number of flagged issues not matching any gold error. We report three recall-style metrics:

- *Average Recall* (AR): $\mathrm{mean}_i \ \frac{|M(r_i)|}{|E_i|}$.
- *Perfect Recovery* (PR): fraction of runs with $|M(r_i)| = |E_i|$.
- *Majority Recovery* (MR): fraction of runs with $|M(r_i)| > \frac{1}{2}|E_i|$.

Each is crossed with two false-positive treatments: *uncorrected* (ignore FPs) and *penalized-disqualifying* (PD: a run counts as failure if $\mathrm{FP}(r_i) > 0$). An exact-set metric combines both:

$$\mathrm{Exact}(r_i) = \mathbb{1}\big[|M(r_i)| = |E_i| \ \wedge \ \mathrm{FP}(r_i) = 0\big]. \quad (1)$$

AR reflects average-case recall for a researcher browsing suggestions; PR and Exact capture whether the model can be trusted to catch everything; MR captures a lenient threshold. The 0-FP gate (PD) models the setting in which the researcher must verify each flag—a noisy reviewer whose flags are wrong as often as right does not, on balance, save the researcher time. We do *not* claim the 0-FP gate is the only correct scoring rule: some non-gold flags may reflect genuine ambiguities in the author's writing. Rather, it is an intentionally conservative proxy for reviewer time cost.

4.3. Qualitative Axes

In parallel, each response is rated by a domain-expert reviewer along three ordinal axes—*helpfulness*, *correctness*, and *alignment*—each on the scale {no issues, minor issues, major issues} mapped to $\{0, 1, 2\}$. Reviewers additionally assign an overall quality rating on a 1–5 Likert scale. We view these as important auxiliary annotations rather than headline metrics: a scientific assistant that identifies some real errors but phrases them vaguely, overstates them, or mismatches the expertise level of the draft may still fail to save researchers time.

5. Results

5.1. Quantitative Findings

Table 3 presents the quantitative results across all scoring regimes. Three findings stand out.

Partial criticism is common, but fully correct review remains rare. Raw recall is meaningfully above zero for all models, which is desirable for a benchmark intended to measure a useful capability rather than impossible puzzle-solving. At the same time, exact-set match under the 0-FP

Table 3. Quantitative results on SCIREVIEW. "Standard" ignores false positives; "0-FP gate" disqualifies any response containing a misidentified error. All numbers in percent; higher is better.

	Standard			0-FP gate		
Model	AR	PR	MR	AR	Exact	$>1/2$
GPT-5.4	62%	6%	71%	10%	0%	12%
Gemini 3.1 Pro	57%	0%	60%	45%	0%	48%
Claude Opus 4.6	46%	0%	33%	24%	0%	19%

Table 4. Qualitative assessment. Helpfulness, correctness, and alignment are scored on $\{0, 1, 2\}$ (0 = no issues, 2 = major issues); lower is better. Overall is a 1–5 Likert scale; higher is better.

Model	**Help.** $\downarrow$	**Corr.** $\downarrow$	**Align.** $\downarrow$	**Overall** $\uparrow$
GPT-5.4	0.60	1.30	0.25	2.65
Gemini 3.1 Pro	0.40	1.30	0.05	3.10
Claude Opus 4.6	0.55	1.55	0.35	2.10

gate is 0% for every model, and standard perfect recovery is 6% for the best model and 0% for the others. This is consistent with SPOT's finding that no frontier model reliably catches all material errors in a scientific document (Son et al., 2025).

Rankings invert under false-positive penalization. GPT-5.4 leads on every uncorrected metric, but once misidentifications disqualify a run, Gemini 3.1 Pro becomes the strongest system—its AR drops only 12 percentage points (from 57% to 45%), compared with GPT-5.4's collapse of 52 points ($62\% \rightarrow 10\%$). The gap reflects a qualitative difference: GPT-5.4 tends to flag more candidate errors, which increases recall when false positives are free but collapses under even a modest precision requirement. In practical research settings, this trade-off matters directly: an assistant that raises many non-errors may waste scientist time even if its nominal recall is high.

The benchmark is not uniformly adversarial. SCIREVIEW contains both baseline and challenging errors across items, with difficulty unevenly distributed. Notably, difficulty estimated during limited author-side pre-screening can shift under broader repeated evaluation: some errors classified as "challenging" during single-shot authoring were found by at least one model across multiple independent runs. This argues for calibrating final difficulty labels against multi-sample evaluation rather than relying only on one-shot checks during construction.

5.2. Qualitative Findings

Table 4 summarizes the expert-rated qualitative assessments.

Correctness is the weakest axis for every model. All three models score between 1.30 and 1.55 on correctness, above the "minor issues" threshold. Reviewers commonly flagged cases where a model identified a real error but misattributed its cause, or added correct-sounding but subtly incorrect technical elaborations.

Gemini 3.1 Pro has the best alignment (0.05) by a wide margin: its responses matched the expert tone of the passage most consistently. Manual inspection suggests Gemini is more likely to respond at the level of rigor of the original, whereas GPT-5.4 and Claude occasionally shift toward tutorial-style explanations inappropriate to a research-proposal context.

Overall quality ranks Gemini $>$ GPT $>$ Claude, inverting the raw-recall ranking. The combined qualitative and penalized-quantitative picture suggests that Gemini's more conservative behavior—fewer flags, fewer mistakes when it does flag—produces responses that expert reviewers find more useful, even when raw recall is lower. This consistency between quantitative precision and qualitative usefulness suggests that the 0-FP gate, though conservative, may track operational value more closely than raw recall alone.

5.3. Representative Failure Modes

SCIREVIEW tests more than factual recall. Table 5 highlights three error types that are locally plausible in context but materially change what a careful researcher should conclude. Full items are provided in the appendix.

Across these cases, a pattern emerges: frontier models fail most consistently on errors that require (a) tracking a latent physical or mathematical parameter through a chain of reasoning, (b) distinguishing levels of analysis (group vs. individual), or (c) noticing when plausible-sounding domain rhetoric is logically disconnected from the evidence presented.

6. Discussion

A distinct gap in AI-scientist evaluation. A scientist who uses an AI assistant does not only need help generating new content. They also need a system that can reliably say, "this step is invalid," "this conclusion does not follow," or "this edge case breaks your argument." Existing benchmarks do not specifically probe this ability. Claim verification and citation attribution isolate narrower scholarship tasks (Wadden et al., 2020; Press et al., 2024). Reproducibility and replication benchmarks test execution and reconstruction (Siegel et al., 2024; Starace et al., 2025). SPOT studies published failures with severe post-publication consequences (Son et al., 2025). SCIREVIEW is aimed earlier in the pipeline: draft-time critique of realistic research-grade scientific writing.

Table 5. Representative error types in SCIREVIEW. Each is locally plausible in context but scientifically consequential if left uncorrected.

Domain	Embedded error	Why it matters
Chemistry	The text places the catalyst on the wrong face (endo vs. exo) of a bicyclic substrate and reports the opposite final enantiomer, $(1S, 6R)$ instead of $(1R, 6S)$.	The stereochemical mistake changes which bioactive product is supported; the writeup ends up justifying the wrong enantiomer for pharmaceutical use. Both errors point the same "wrong" direction, so the passage remains internally consistent. Catching it requires tracking 3D geometry through a multi-step reaction.
Neuroscience	The text claims that group-level neuroimaging findings can be deployed as validated *individual* diagnostic tools.	The error confuses population-level association with individual-level diagnosis—a clinically consequential form of overreach. Models pretrained on grant rhetoric are biased to accept such claims as normal "Clinical Implications" prose, even when the body of the text only supports a group-level claim.
Math. physics	The text states that Price's law extends to the *full* Reissner–Nordström black-hole family, including the extremal case.	The claim silently fails at extremality, where the Aretakis instability changes horizon behavior (Aretakis, 2015). Detecting the mistake requires understanding how vanishing surface gravity breaks the generic sub-extremal argument.

Why exact-set match matters. For scientific review, partial credit is not the whole story. Missing one crucial error may invalidate an entire research direction, while adding several non-errors creates costly distraction. This is why we view exact-set match and false-positive-aware variants as core metrics rather than optional extras. A benchmark for AI scientists should reward noticing the *right* problems.

Why expert-authored errors matter. Error-injection benchmarks that use LLMs to insert errors (FLAWS, PaperAudit-Bench) inherit the inserter model's error distribution; this is valuable for scale but creates a blind spot where the inserter and the detector share biases. Retraction-based benchmarks (SPOT) capture real errors but sample disproportionately from high-severity cases. The expert-authored protocol used here produces errors drawn from the distribution a domain expert would *make*—including the subtle, plausibility-preserving errors that are most operationally costly and most likely to survive peer review.

Workflow realism. A useful benchmark for AI scientists should ask whether a model's critique would be worth acting on in practice. This is why SCIREVIEW tracks auxiliary qualities such as helpfulness, correctness-as-written, and alignment to expertise level. They point toward a broader evaluation agenda in which scientific-review assistants are judged both on what they catch and on how productively they communicate it.

7. Future Work

We note several important directions for future work:

Adversarial filter is model-specific. "Challenging" errors are defined relative to the three frontier models used at construction time, and as our difficulty calibration analysis shows (Section 5), these labels can shift under repeated evaluation. Future iterations will refresh the filter panel periodically and publish difficulty flags independently so researchers can recompute them.

Domain balance. The current benchmark shows useful breadth, but future versions should report domain composition explicitly and guard against over-representing fields whose errors are especially easy to phrase in prose.

Static, non-interactive evaluation. The benchmark currently evaluates stand-alone reviews of a fixed text. Many real deployments will be interactive: a model may ask clarifying questions, request references, or revise its critique after feedback. Extending SCIREVIEW to collaborative review settings is an important next step.

8. Conclusion

We introduced SCIREVIEW, a benchmark for evaluating AI scientists as scientific reviewers. The benchmark targets a capability central to trustworthy AI-assisted research but under-measured by current evaluations: detecting realistic, high-consequence conceptual errors embedded in scientific writeups before publication. Our design emphasizes expert-authored texts, natural error injection, adversarial difficulty calibration, and exclusion of trivial artifacts. Results show that frontier models can often identify isolated problems but still struggle to produce complete, precise scientific reviews: no model exceeds 6% perfect recovery, and the strongest uncorrected recaller collapses from 62% to 10% average recall under a 0-FP gate. We hope SCIREVIEW

helps move AI-for-science evaluation beyond generation and execution toward the harder question of whether AI systems can reliably improve scientific rigor.

Broader Impact

This paper presents work whose goal is to advance the field of Machine Learning Research by providing a more rigorous evaluation framework for LLM-based scientific assistants. The benchmark is intended to improve the reliability of AI systems used in scientific workflows. We have discussed relevant risks (over-reliance, contamination, overfitting to error types) in Section 7 and explicitly recommend that frontier models not be deployed as unsupervised research auditors on the basis of SCIREVIEW performance alone.

References

Hongwei Liu, Junnan Liu, Shudong Liu et al. ATLAS: A high-difficulty, multidisciplinary benchmark for frontier scientific reasoning. *arXiv preprint arXiv:2511.14366*, 2025.

Stefanos Aretakis. Horizon instability of extremal black holes. *Advances in Theoretical and Mathematical Physics*, 19(3):507–530, 2015.

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.

Howard Bauchner and Frederick P. Rivara. Use of artificial intelligence and the future of peer review. *Health Affairs Scholar*, 2(5), 2024.

C. Glenn Begley and Lee M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

Center for AI Safety, Scale AI, and the HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026.

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

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

Timothy M. Errington, Maya Mathur, Courtney K. Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A. Nosek. Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601, 2021.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2024.

John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8):e124, 2005.

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

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 2024b.

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.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. LLM4SR: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, et al. DiscoveryBench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, Washington, DC, 2019.

Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong, and Shafiq Joty. Hard2Verify: A step-level verification benchmark for open-ended frontier math. *arXiv preprint arXiv:2510.13744*, 2025.

Russell A. Poldrack, Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18:115–126, 2017.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. CiteME: Can language models accurately cite scientific claims? *arXiv preprint arXiv:2407.12861*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

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

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, et al. When AI co-scientists fail: SPOT—a benchmark for automated verification of scientific research. *arXiv preprint arXiv:2505.11855*, 2025.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, et al. SciCode: A research coding benchmark curated by scientists. *arXiv preprint arXiv:2407.13168*, 2024.

Songjun Tu et al. PaperAudit-Bench: Benchmarking error detection in research papers for critical automated peer review. *arXiv preprint arXiv:2601.19916*, 2026.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of EMNLP*, pp. 7534–7550, 2020.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. *arXiv preprint arXiv:2210.13777*, 2022.

Sarina Xi, Vishakh Rao, Justin Payan, and Nihar B. Shah. FLAWS: A benchmark for error identification and localization in scientific papers. *arXiv preprint arXiv:2511.21843*, 2025.

Bingyang Ye, Shan Chen, Jingxuan Tu, Chen Liu, Zidi Xiong, Samuel Schmidgall, and Danielle S. Bitterman. Proof of Time: A benchmark for evaluating scientific idea judgments. *arXiv preprint arXiv:2601.07606*, 2026.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Zekun Zhou, Xiaocheng Feng, Lei Huang, et al. From hypothesis to publication: A comprehensive survey of AI-driven research support systems. *arXiv preprint arXiv:2503.01424*, 2025.

A. Additional Scoring Details

A.1. Penalized-Canceling Scoring

For completeness, we report a third false-positive treatment: *penalized-canceling* (PC), where the count of correctly identified errors is reduced by the number of misidentified errors, floored at zero.

Table 6. Penalized-canceling results. Higher is better. All numbers in percent.

Metric	GPT-5.4	Gemini 3.1 Pro	Claude Opus 4.6
AR–PC	29%	51%	33%
PR–PC	0%	0%	0%
MR–PC	17%	52%	19%

The PC ranking agrees qualitatively with the PD ranking (Gemini best, Claude worst), but the semantics of "subtracting" one error count from another are ambiguous: a run with four correct identifications and four misidentifications is not obviously equivalent to a run with zero of each. We therefore emphasize PD in the main text and include PC for readers interested in rank-stability.

A.2. Tolerance Variant

One additional variant considered was allowing up to two misidentified errors before penalization. This was not adopted because it conflates error-identification quality with writing-quality feedback: model-flagged "errors" that are not expert-listed may still surface genuine ambiguities in the author's writing. We plan to address writing-quality feedback as a separate benchmark axis.

B. Full Example: Mathematical Physics

Passage excerpt. "Let us now turn the charge on and consider the full Reissner–Nordström (RN) family which admits a black hole (we take $Q > 0$ and $a = 0$). To further simplify the discussion, we will consider from now onwards that the initial data on N_1 are in fact compactly supported (and hence trivially conformal). Aretakis and collaborators rigorously showed in 2016 that Price's law extends to these spacetimes."

Expert-annotated error. The statement that Price's law applies to the full Reissner–Nordström family is wrong because this family contains the *extremal* case ($|Q| = M$, surface gravity $\kappa = 0$), for which the decay behavior along the event horizon differs qualitatively—this is the Aretakis instability. Price's law in the form stated applies to sub-extremal RN; in the extremal case, transverse derivatives of a linear scalar field fail to decay along the horizon. This error overlooks a major discovery in black-hole dynamics.

Why the error is subtle. In the sub-extremal regime, positive surface gravity produces a red-shift effect at the horizon that stabilizes perturbations. A reader who does not track the surface-gravity parameter through the phrase "full RN family" can miss that the claim extrapolates sub-extremal intuition into the extremal regime, where the stabilizing mechanism disappears.

C. Full Example: Chemistry

Passage context. The passage discusses synthetic methods for accessing enantiomerically enriched β-amino acids used in bioactive peptide synthesis, where metabolic stability, potency, and safety profiles are highly enantiomer-selective.

Errors.

1. "...catalyst resided on the **endo** face of the bicyclic substrate, despite the cyclohexane portion being oriented toward the sterically cumbersome carbocyclic portion of the succinimide."
   *Correct: the catalyst resides on the **exo** face.*
2. "...delivering methyl **(1S,6R)**-6-((methoxycarbonyl)amino)cyclohex-3-ene-1-carboxylate."
   *Correct: the product is the **(1R,6S)** enantiomer.*

Why the errors are subtle. The face assignment and absolute configuration together determine the enantiomer of the final

product; both errors point the same "wrong" direction, so the passage remains internally consistent and the downstream chemistry description is self-coherent. Catching this requires tracking the three-dimensional geometry of the reaction and re-deriving stereochemical labels rather than reading them as asserted.

D. Full Example: Neuroscience

Passage context. A study on brain plasticity in patients with painful diabetic peripheral neuropathy (PDN) reports group-level neuroimaging differences between PDN patients and controls.

Error. "These findings highlight the importance of deploying neuroimaging biomarkers as validated individual diagnostic tools in PDN diagnosis and monitoring."

Why the error is subtle. The phrase appears in a "Clinical Implications" subsection, where speculative, forward-looking claims are grammatically and stylistically expected. In grant applications, abstracts, and promotional writing, such claims are common and often tolerated; models pretrained on this genre may be biased to read them as part of the normal register rather than as logical overreach. The group-to-individual inference leap is a long-standing issue in neuroimaging methodology (Poldrack et al., 2017) and has direct patient-safety consequences if accepted at face value.