PAPEREVALUATION: BENCHMARKING AGENT-GENERATED SCIENTIFIC PAPERS WITH EXECUTION EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly capable of performing complex, long-horizon research tasks. However, existing scientific benchmarks provide limited coverage of end-to-end workflows from data analysis and experimental validation to complete manuscript generation, and lack targeted evaluation that connects research execution to the final manuscript. We introduce PaperEvaluation, a benchmark for evaluating agent-generated scientific papers using execution evidence. The benchmark comprises 80 research task packages derived from published papers across 13 research domains. Each task provides a research question, the necessary data, and a controlled execution environment while withholding the source paper and its conclusions. Agents must independently conduct the research and produce a complete manuscript. To evaluate this long-horizon research process, we propose R3Eval, a framework with three dimensions: Research Completeness, Reporting Rigor, and Reasoning Alignment. Research Completeness assesses whether the required scientific work is supported by execution evidence. Reporting Rigor examines whether the manuscript accurately reports its methods and findings. Reasoning Alignment compares the generated and source papers for functional alignment of their scientific arguments. Finally, we evaluate 14 agent systems based on PaperEvaluation, revealing distinct performance profiles across these dimensions through an evaluation of long-horizon research that connects execution to the final manuscript.

1 INTRODUCTION

Advances in large language models (LLMs) and Harness systems have extended agents from simple interactions to complex tasks requiring long-horizon coordination, creating new evaluation challenges. Recent systems support structured hypothesis generation and experimental validation (Gottweis et al., 2026), autonomous experimentation and manuscript production (Yamada et al., 2025), and scientific workflows in chemistry and biomedicine (Li et al., 2026; Huang et al., 2025; Bran et al., 2024). Scientific research offers a suitable setting for assessing long-horizon capabilities: agents must turn research questions into scientific arguments grounded in analysis and validation while accurately reporting their work.

Existing scientific benchmarks cover problem solving, computational reproducibility, and innovative research (Chen et al., 2025b; Siegel et al., 2025; Wu et al., 2026), but often emphasize individual problems, code implementation, or result comparison. Complete research-to-manuscript workflows and their integrated evaluation remain underexplored. We argue that an end-to-end research benchmark should provide executable task packages from published papers across disciplines, supporting agents from research questions to complete manuscripts. Comprehensive evaluation should link execution to manuscripts, assessing the work performed and its support for reported conclusions rather than results or manuscript quality alone.

To meet these requirements, we introduce PaperEvaluation, a benchmark that links agent-generated scientific papers to the execution evidence supporting their claims. We construct 80 tasks from published papers across 13 research domains. Each task provides a research question and frozen data within a standardized execution environment. The source paper and derived evaluation criteria remain

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

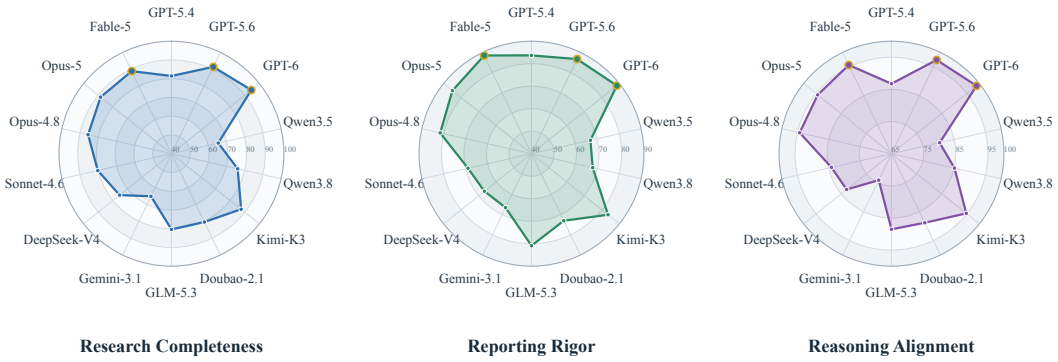

Figure 1: Results of the three complementary evaluation dimensions in PaperEvaluation: Research Completeness, Reporting Rigor, and Reasoning Alignment.

hidden, so the agent must choose an approach and develop its own analysis rather than organize the work around a supplied answer. We collect execution traces, analytical artifacts, and final manuscripts from this environment to evaluate agents' long-horizon research capabilities.

To evaluate the work performed and the resulting scientific argument using these materials, we propose R3Eval, a three-dimensional evaluation framework comprising Research Completeness, Reporting Rigor, and Reasoning Alignment. Specifically, Research Completeness uses execution evidence to determine whether the agent has completed the applicable scientific work. Reporting Rigor examines whether the manuscript accurately reports and supports its claims. Reasoning Alignment compares its argument with the source study to assess whether it answers the same research question through a functionally comparable reasoning process. By linking execution evidence to the manuscript's reporting and scientific argument, R3Eval assesses long-horizon research capabilities in terms of research completeness, reporting rigor, and functional alignment of scientific reasoning, rather than inferring research quality from the final manuscript alone.

Finally, for a common basis of comparison, we introduce the EvidenceFirst Scaffold, which guides research planning and manuscript preparation without prescribing an analytical method. All evaluated systems use the same Scaffold and interact with the environment through their fixed model–Harness configurations, with R3Eval evaluating their outputs after submission.

Our contributions are:

- We construct task packages from 80 published papers across 13 research domains, providing a benchmark for evaluating agents on long-horizon research tasks.
- We provide the EvidenceFirst Scaffold for research planning and manuscript writing and propose R3Eval to assess agents' long-horizon research capabilities across three dimensions using execution evidence and final manuscripts.
- We compare 14 fixed Model–Harness configurations and analyze their performance profiles and characteristic failure modes, with an overview in Fig. 1.

2 Related Work

2.1 Benchmarks for Scientific Research

As LLMs advance, agent research spans GAIA's grounded reasoning and tool-use evaluation (Mialon et al., 2024), ReAct's interleaved reasoning and environment actions (Yao et al., 2023), SWE-agent's interface-based improvements to long software workflows (Yang et al., 2024), and reliability measurements over longer tasks (Kwa et al., 2025). Scientific research supports open-ended, long-horizon evaluation through multi-step analysis and validation. Benchmarks cover executable data science, DataSciBench and InfiAgent-DABench (Zhang et al., 2026; Hu et al., 2024), multi-step dataset-based discovery, DiscoveryBench (Majumder et al., 2025), iterative ML experimentation and engineering, MLAgentBench and MLE-bench (Huang et al., 2024; Chan et al., 2025), and

Benchmark	Objective	Tasks	Deliverable	Evaluation
PaperBench	Result replication	20	Code	Execution + rubrics
MLR-Bench	Open-ended discovery	201	Paper	LLM rubrics
InnovatorBench	Method optimization	20	Code	Hidden metrics
NatureBench	Method optimization	90	Code	SOTA-relative metrics
Ours	Question-to-paper	80	Paper	Evidence-chain rubrics

Table 1: Comparison of research-agent benchmarks (Starace et al., 2025; Chen et al., 2025a; Wu et al., 2026; Wang et al., 2026). Counts refer to full benchmark task sets, not necessarily the evaluated subsets; deliverables and evaluation are abbreviated. PaperEvaluation jointly examines traces, artifacts, and manuscripts within the visible task scope.

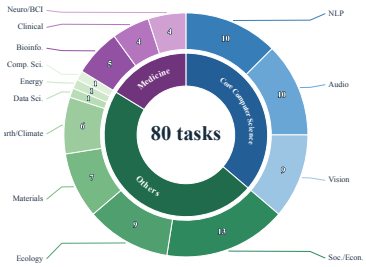

Figure 2: Distribution of the 80 benchmark tasks across 13 subfields.

broader scientific activities with controlled tools, AstaBench (Bragg et al., 2026). Independent KOSMOS evaluation motivates checking findings against appropriate baselines (Nusrat & Nusrat, 2025). PaperEvaluation instead hides source papers, evaluating research from questions and data and linking manuscripts to evidence of performed work. In Table 1, we show the differences between our benchmark and other research benchmarks.

2.2 Agent Evaluation and Process Rewards

Evaluation increasingly combines final outputs with structured evidence. FActScore decomposes long-form generations into atomic factual claims (Min et al., 2023), while G-Eval uses structured LLM evaluation for generated text (Liu et al., 2023). Studies of LLM judges examine agreement with human preferences and robustness on challenging factual and reasoning comparisons (Zheng et al., 2023; Tan et al., 2025). Process supervision evaluates intermediate reasoning rather than only final outcomes (Lightman et al., 2023); PRMBench and ProcessBench further test fine-grained detection of errors within reasoning processes (Song et al., 2025; Zheng et al., 2025). R3Eval instead uses research-process evidence to assess the scientific argument in a completed paper, separating completion of the required work from accurate reporting and source-referenced reasoning alignment.

3 PaperEvaluation

3.1 Task Formulation

In our benchmark, each task is represented as $\tau = (V, H)$, with visible inputs V and hidden evaluation information H. The visible component is $V = (Q, D, \mathcal{E}, \mathcal{O})$, where Q specifies the research question and necessary scientific constraints, D is the frozen data, $\mathcal{E}$ is the execution environment, and $\mathcal{O}$ is the Submission Contract specifying required deliverables and their format. The hidden component is $H = (M_{\text{src}}, \mathcal{Y})$, where M_{src} is the source paper and $\mathcal{Y}$ contains the evaluation criteria frozen before candidate scoring. These criteria specify the required scientific work and the source study's reasoning structure. Given only V, the agent system $\mathcal{A}$ produces

$$(T, F, M) = \mathcal{A}(V), \tag{1}$$

where T records tool calls, their results, and their order; F contains analytical artifacts, including machine-readable results and validation records; and M is the final manuscript with its figures and tables.

3.2 Benchmark Construction

Paper selection. Published papers provide research questions and scientific arguments that have undergone peer review. Thus, we first collect 2,000 papers across disciplines and screen them for task executability, considering data availability, implementation support, computational feasibility, publication completeness, clarity of evidence and metrics, and task suitability. Using these criteria, we retain papers that can be converted into executable research tasks with the available data and

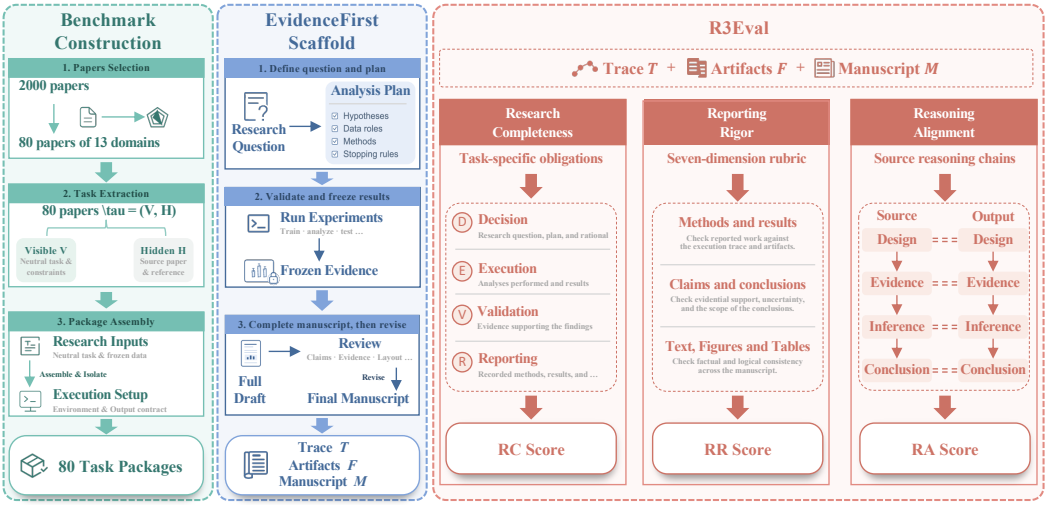

Figure 3: Overview of PaperEvaluation: benchmark construction produces standardized research tasks, the shared EvidenceFirst Scaffold guides the end-to-end workflow from planning to manuscript generation, and R3Eval separately scores Research Completeness, Reporting Rigor, and Reasoning Alignment from the frozen trace, analytical artifacts, and manuscript.

computational resources. We then select 13 research domains and a total of 80 papers from these domains to construct the benchmark. Their distribution across domains is shown in Fig. 2.

Task extraction and packaging. Following paper selection, we use an LLM to convert each paper into a self-contained research task and assemble a corresponding task package. Each package retains the research objective and necessary scientific constraints and provides the research inputs and execution conditions, while withholding the source paper, its conclusions and key results, and the evaluation criteria. The agent chooses its own implementation and develops evidence from the research question and data rather than working backward from a reference answer. We also verify task inputs and record data availability and known reproduction constraints so that evaluation can distinguish unavailable resources from feasible work the agent could have completed but did not.

3.3 EvidenceFirst Scaffold

We use the same EvidenceFirst Scaffold across systems to provide a common basis for comparison. Beyond basic writing requirements, we specify the following three principles to make the research process more traceable and support subsequent evaluation.

Define the research question and plan before execution. To keep the research objective stable and limit result-driven changes to the research question and rationale, the agent clarifies the task's research question and records its analysis plan and stopping rules before execution. The plan must preserve task-required mechanisms and specify the evidence needed to answer the question. Different defensible analytical choices can yield substantially different results (Silberzahn et al., 2018). Recording the plan in advance provides a reference for tracing subsequent research decisions and is intended to reduce time spent repeatedly changing research direction.

Validate and freeze results before writing. To ground manuscript conclusions in validated analytical results, the agent separates analysis selection from final validation to avoid repeatedly adjusting the analysis based on validation outcomes. This separation is intended to reduce threats to research validity, including data leakage (Kapoor & Narayanan, 2022). Before using results in the manuscript, the agent freezes them and their supporting analysis records. Subsequent text, figures, tables, and reported values draw on these records to keep the manuscript consistent with the analyses performed (Kapoor et al., 2023).

Complete the full manuscript before focused revision. Once sufficient evidence is available to support a draft, the agent completes and compiles the full manuscript before conducting a structured review and focused revision. The review checks whether the manuscript's scientific argument is supported by the available evidence and identifies any necessary additional analyses or figures. This ordering is intended to reserve time for delivering a complete manuscript and avoid prolonged cycles of local revision or supplementary analysis.

3.4 EVALUATION

To distinguish completed scientific work from the quality of its presentation and reasoning, we propose R3Eval to evaluate the frozen trace T, artifacts F, and manuscript M along three dimensions:

$$\mathbf{S} = \Phi(T, F, M; V, H) = (S_{\text{comp}}, S_{\text{rep}}, S_{\text{align}}). \tag{2}$$

Here, Φ denotes the evaluator, and the three scores measure Research Completeness (RC), Reporting Rigor (RR), and Reasoning Alignment (RA), respectively. Although the three dimensions share the same input materials, they assess different aspects of the submission: completion of the required work, rigor of its reporting, and functional alignment of its scientific argument with the source study.

3.4.1 RESEARCH COMPLETENESS

A manuscript summarizes the research process but cannot establish on its own whether the required analyses were performed and validated. We therefore propose Research Completeness (RC) to measure the overall extent of research completion. RC combines the execution trace, analytical artifacts, and manuscript to assess *research obligations*, defined as the scientific work required to answer the visible research question together with the evidence needed to establish its completion.

Let $\mathcal{B}_\tau$ be the obligations in the Evaluation Obligation Contract, frozen before candidate scoring. For each obligation o, we assess evidence of the research decision, execution, validation, and reporting. The corresponding credits form $\mathbf{s}_o = (s_{o,d}, s_{o,e}, s_{o,v}, s_{o,r})^\top$, with each credit in $\{0, 0.5, 1\}$ for absent, partial, or sufficient evidence. A normalized stage-weight vector $\boldsymbol{\lambda}$ combines these credits. To prevent strong planning or reporting from compensating for missing execution evidence, we constrain the result by an upper bound κ, determined by the obligation state z_o and supporting-evidence conditions e_o:

$$\widetilde{c}_o = \boldsymbol{\lambda}^\top \mathbf{s}_o, \qquad c_o = \min\{\widetilde{c}_o, \kappa(z_o, e_o)\}\,. \tag{3}$$

We then aggregate obligations that apply to the visible task. Let $\mathcal{B}_\tau^+ \subseteq \mathcal{B}_\tau$ denote this eligible set and v_o the pre-specified importance of obligation o. Available but unattempted work remains eligible with zero credit; genuinely unavailable requirements are excluded. Research Completeness is

$$S_{\text{comp}} = 100\,\frac{\sum_{o \in \mathcal{B}_\tau^+} v_o c_o}{\sum_{o \in \mathcal{B}_\tau^+} v_o}. \tag{4}$$

In this way, RC measures the completion of required research work and helps identify manuscript conclusions unsupported by analysis and validation.

3.4.2 REPORTING RIGOR

Completing a study requires not only conducting the necessary analyses and validation but also reporting accurate, rigorous methods and findings. We introduce Reporting Rigor (RR), using a single structured LLM Judge to assess manuscript M across seven dimensions $\mathcal{D}_{\text{rep}}$: research framing, methodological reporting, evidence support, statistical reasoning, claim calibration, logical coherence, and visual integrity. Unlike RC, RR assesses whether the manuscript faithfully reports the work performed and calibrates its conclusions to the available evidence. Specifically, let $\widehat{q}_d \in [0, 1]$ be the normalized Judge score for dimension d, and β_d its pre-specified weight. Reporting Rigor is

$$S_{\text{rep}} = 100 \sum_{d \in \mathcal{D}_{\text{rep}}} \beta_d \widehat{q}_d, \qquad \sum_{d \in \mathcal{D}_{\text{rep}}} \beta_d = 1. \tag{5}$$

The rubric distinguishes accurately reported research from factual inconsistencies and conclusions that exceed the available evidence, even when the underlying task has been completed.

3.4.3 Reasoning Alignment

Finally, we introduce Reasoning Alignment (RA) to examine differences in scientific reasoning between the generated study and the source paper. We first extract task-specific reasoning chains from the source paper and freeze them before evaluation. These chains guide the comparison between the generated manuscript and the source paper to assess whether their scientific arguments are functionally aligned. Each chain links the research question to a conclusion through the study design and evidence, allowing the Judge to compare the roles of individual components and the relations supporting the inference. We set $\mathcal{C}$ to the set of source reasoning chains. For chain c, let $\mathcal{K}_c^+$ denote its eligible components and $\mathcal{L}_c$ the argumentative relations between them. The Judge assigns component credits $r_{c,k}$ and relation credits $r_{c,\ell}$, giving the chain-level score

$$a_c = \eta \frac{\sum_{k \in \mathcal{K}_c^+} \alpha_k r_{c,k}}{\sum_{k \in \mathcal{K}_c^+} \alpha_k} + (1-\eta) \frac{1}{|\mathcal{L}_c|} \sum_{\ell \in \mathcal{L}_c} r_{c,\ell}, \tag{6}$$

where α_k weights components and η controls the balance between component and relation scores. We aggregate chains using their pre-specified weights w_c:

$$S_{\text{align}}^{\text{raw}} = 100 \frac{\sum_{c \in \mathcal{C}} w_c a_c}{\sum_{c \in \mathcal{C}} w_c}, \qquad S_{\text{align}} = \min\left\{ S_{\text{align}}^{\text{raw}}, \kappa_{\text{crit}} \right\}. \tag{7}$$

Here, κ_{crit} limits the score when a critical reasoning chain fails, preventing other aligned components from compensating for a central inferential defect. It equals 100 when no such failure occurs. Thus, RA assesses whether the generated paper develops an evidence-supported argument for the same question while allowing alternative scientific approaches.

Further details on paper selection, task package construction, and R3Eval are provided in the appendix.

4 Experiments

4.1 Experimental Setup

We evaluate the 14 agent systems on the 80 PaperEvaluation tasks across 13 research domains. Each system is a fixed Model–Harness configuration using the same EvidenceFirst Scaffold. Claude Code directly runs Claude-Opus-5 (Anthropic, 2026c), Claude-Fable-5 (Anthropic, 2026a), Claude-Opus-4.8 (Anthropic, 2026b), Claude-Sonnet-4.6 (Anthropic, 2026d), and Kimi-K3 (Kimi Team, 2026); DeepSeek-V4 (DeepSeek-AI, 2026), GLM-5.3 (GLM-5 Team, 2026; Z.ai, 2026), Doubao-Seed-2.1 (ByteDance Seed, 2026), Qwen3.8 (Qwen Team, 2026b), and Qwen3.5-397B-A17B (Qwen Team, 2026a) use protocol conversion. Codex CLI runs GPT-5.4 (OpenAI, 2026a), GPT-5.6 Sol (OpenAI, 2026b), and GPT-6 (OpenAI, 2026c) through the Responses interface with high reasoning effort. Gemini CLI runs Gemini-3.1-Pro (Google DeepMind, 2026).

All model-assisted steps in benchmark construction and evaluation use GPT-5.4, including offline evidence extraction and a single balanced structured review supplying the primary RC, RR, and RA judgments.

Additional experimental and evaluation details appear in the appendix.

4.2 Main Results

We begin with the main evaluation results for the 14 agent systems, as shown in Table 2. GPT-6, GPT-5.6 Sol, and Claude-Fable-5 form the leading group, with GPT-6 achieving the highest score on all three dimensions. GPT-5.6 Sol ranks second in research completeness and reasoning alignment

System	Research Completeness		Reporting Rigor		Reasoning Alignment	
	Score ↑	A-ICC ↑	Score ↑	A-ICC ↑	Score ↑	A-ICC ↑
GPT-6	**94.46**	0.956	**88.53**	0.983	**98.86**	0.957
GPT-5.6 Sol	<u>91.38</u>	0.835	86.67	0.850	<u>97.38</u>	0.948
Claude-Fable-5	88.86	0.895	<u>88.44</u>	0.920	95.63	0.949
Claude-Opus-5	88.38	0.810	84.95	0.858	94.36	0.983
Kimi-K3	87.51	0.910	83.52	0.837	94.77	0.854
Claude-Opus-4.8	85.62	0.960	81.62	0.845	94.30	0.930
Doubao-Seed-2.1	80.43	0.852	73.06	0.829	88.85	0.887
GLM-5.3	80.37	0.899	80.88	0.973	88.52	0.878
GPT-5.4	81.45	0.790	83.70	0.848	86.77	0.888
Claude-Sonnet-4.6	80.31	0.871	68.97	0.767	84.15	0.930
Qwen3.8	76.23	0.612	67.91	0.683	85.10	0.762
DeepSeek-V4	75.39	0.906	66.75	0.807	82.84	0.786
Gemini-3.1-Pro	65.21	0.916	66.60	0.757	74.14	0.704
Qwen3.5	65.58	0.730	66.87	0.948	80.36	0.825
Overall J-ICC ↑	0.908		0.940		0.877	
Expert ICC ↑	0.912		0.935		0.883	

Table 2: Performance and reliability across RC, RR, and RA. **Bold** and <u>underlined</u> values indicate the best and second-best performance, respectively.

but third in reporting rigor, where Claude-Fable-5 ranks second. The different rankings of GPT-5.6 Sol and Claude-Fable-5 show that stronger completion and reasoning alignment scores do not necessarily imply more rigorous reporting. These differences extend beyond the leading group. Kimi-K3 outperforms Claude-Opus-5 in RA despite scoring lower in RC and RR. Reporting the three scores separately reveals differences in system performance that would be missed by considering completion or reasoning alignment alone.

The same distinction appears among systems with similar levels of completeness. Doubao-Seed-2.1 and GLM-5.3 have nearly identical RC scores, yet GLM-5.3 performs better in RR. Similar aggregate completeness can thus coexist with differences in how rigorously the work is reported. Reasoning alignment provides another distinction: Qwen3.8 ranks below Claude-Sonnet-4.6 in RC and RR but above it in RA. At the lower end, Gemini-3.1-Pro and Qwen3.5 rank near the bottom on every axis. Together, these profiles distinguish systems that lag across dimensions from those whose relative strengths depend on the aspect of research being evaluated.

4.3 Stability Analysis

Because agent-generated manuscripts may vary across independent runs, we assess stability at two levels: task scores across independent agent runs and Judge scores across repeated evaluations of a fixed submission. Both are measured using the intraclass correlation coefficient (ICC):

$$\mathrm{ICC} = \frac{\sigma^2_{\mathrm{object}}}{\sigma^2_{\mathrm{total}}}, \tag{8}$$

where $\sigma^2_{\mathrm{object}}$ is between-object variance and $\sigma^2_{\mathrm{total}}$ is total variance under the specified ICC model. J-ICC treats frozen submissions as objects and fresh Judge evaluations as repeated measurements. A-ICC treats tasks within a system as objects and independent agent runs as repeated measurements. Continuation rounds belong to one run.

J-ICC measures whether the same submission receives consistent evaluations. Table 2 shows high scoring repeatability across all three dimensions, with RR highest and RA comparatively lower. The latter may reflect differences in judging reasoning details when comparing arguments across papers. The generally high J-ICC values provide evidence of evaluator consistency that supports the interpretation of A-ICC.

Candidate evidence	RC MAE ↓	RR MAE ↓	RA MAE ↓
Manuscript (M)	15.3	10.2	8.1
Manuscript + artifacts (M, F)	7.1	6.7	7.8
Manuscript + artifacts + trace (M, F, T)	**3.4**	**4.8**	**6.2**

Table 3: Ablation study on the value of execution evidence.

Perturbation	ΔRC	ΔRR	ΔRA	Perturbation	ΔRC	ΔRR	ΔRA
Language polishing	0.20	0.40	0.30	Missing execution	13.07	2.00	5.59
Missing validation	10.08	4.52	2.59	Numerical inconsistency	0.14	13.00	2.09
Unsupported claim	0.21	8.20	15.09	Incorrect rationale	0.32	3.90	18.59

Table 4: Ablation study on the controlled error injection.

A-ICC measures whether systems perform consistently across independent runs. The mean A-ICC across systems and dimensions is 0.860, indicating generally good run consistency under fixed data and common execution settings. Stability nevertheless varies across systems. For example, Claude-Opus-4.8 has a lower mean RC score than GPT-5.6 Sol but a higher RC A-ICC, indicating more consistent performance across runs on this dimension.

4.4 Agreement with Expert Evaluators

To examine whether R3Eval captures the specialized judgments involved in scientific evaluation, we invited five experts with extensive experience in submitting scientific papers to assess submissions across RC, RR, and RA. For RR, the experts assess the manuscript's reporting accuracy and support for its claims. For RC, structured evidence derived from the original trace enables them to examine whether the required work was performed and whether its results support the manuscript. For RA, they compare the generated and source papers for functional alignment of their scientific arguments. Table 2 reports ICCs between automated scores and expert assessments of 0.912 for RC, 0.935 for RR, and 0.883 for RA. This close agreement supports R3Eval as an assessment of research completion, reporting rigor, and reasoning alignment that is consistent with expert judgment within the evaluated sample.

4.5 The Value of Execution Evidence

To determine whether access to the underlying research process improves agreement with expert assessments, we conduct an ablation study that varies the candidate evidence available to the GPT Judge while keeping the submissions and expert reference scores fixed. We compare manuscript-only evaluation (M), evaluation with additional analytical artifacts (M, F), and the standard R3Eval setting with both artifacts and the execution trace (M, F, T). Task materials, evaluation contracts, source reasoning rubrics, and Judge settings remain unchanged across conditions.

We report the mean absolute error (MAE) for each dimension in Table 3. We can observe that providing execution evidence reduces MAE across all three dimensions, with the largest reduction in RC, from 15.3 with the manuscript alone to 3.4 with full evidence. Adding analytical artifacts accounts for much of the improvement in RC and RR, consistent with their value for checking whether reported work and results are supported by the submission. Access to the trace further reduces error in every dimension. RA changes little when artifacts alone are added but improves more when the trace is also available, suggesting that information about the research process helps assess the scientific argument beyond the manuscript and source rubric.

4.6 Controlled Error Injection

To examine how R3Eval responds to changes in expression and scientific content, we apply six controlled perturbations to submissions. Language polishing changes expression while aiming to preserve scientific content, whereas missing execution and missing validation remove evidence of the corresponding work. Numerical inconsistency, unsupported claims, and incorrect rationale introduce

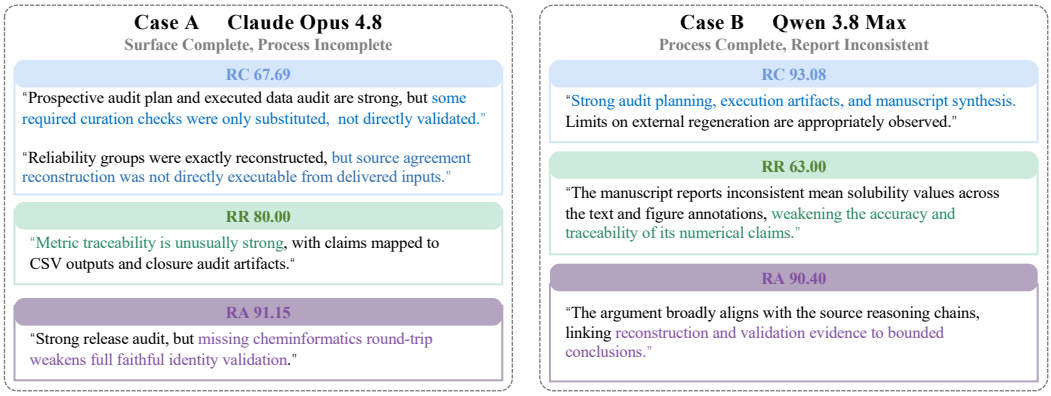

Figure 4: Two same-task cases illustrate different score profiles and evidence issues.

defects in reporting or inference. Table 4 reports the mean score drop, defined as the original score minus the perturbed score.

Language polishing produces only small mean changes, whereas removing execution or validation evidence produces the largest drops in RC. Numerical inconsistency primarily affects RR, reflecting its focus on accurate reporting. Incorrect rationale has the largest effect on RA, which evaluates how evidence supports the scientific argument. Unsupported claims reduce both RR and RA: they weaken the manuscript's reporting and the inference drawn from its evidence. These patterns illustrate differential sensitivity to scientific defects while showing relatively little sensitivity to the tested changes in expression.

4.7 Case Studies

We present a case study to illustrate how R3Eval distinguishes different aspects of research performance. Fig. 4 compares submissions from two systems on the same task and shows the evidence underlying their assessments.

In Case A, Claude Opus 4.8 performs well on RR and RA but achieves a comparatively lower RC assessment. Its trace records alternative analyses and source-resource limitations, while some applicable research obligations remain incompletely supported. This case shows that strong reporting and reasoning alignment can coexist with incomplete research execution. Since RC excludes obligations adjudicated as unavailable, the lower RC reflects gaps in support for applicable work rather than a direct penalty for resource limitations. In contrast, Qwen 3.8 Max performs well on RC and RA. Its trace and artifacts support core audits, group reconstruction, and predictive analysis. However, the manuscript inconsistently reports mean and median LogS across the text, figure annotations, and captions. This discrepancy illustrates why RR is needed alongside RC, as evidence that the research was performed does not establish that its findings were accurately reported.

5 Conclusion

We introduce PaperEvaluation to evaluate agent-generated scientific papers together with the execution evidence supporting their claims. The benchmark comprises 80 published-paper tasks across 13 domains, and its three-axis R3Eval framework connects execution evidence to reporting accuracy and scientific reasoning. Evaluation of 14 systems reveals distinct performance profiles, while repeated assessments, expert comparisons, and controlled perturbations support the evaluation's repeatability and diagnostic value in the studied setting. This approach makes research-agent assessment more informative than judging the final paper alone.

AI use disclosure. We used AI tools for manuscript polishing, literature retrieval and discovery, and synthetic data generation.

REFERENCES

Anthropic. System Card: Claude Fable 5 & Claude Mythos 5. System card, Anthropic, June 2026a. URL https://www.anthropic.com/claude-fable-5-mythos-5-system-card.

Anthropic. System Card: Claude Opus 4.8. System card, Anthropic, May 2026b. URL https://www.anthropic.com/claude-opus-4-8-system-card.

Anthropic. System Card: Claude Opus 5. System card, Anthropic, July 2026c. URL https://www.anthropic.com/claude-opus-5-system-card.

Anthropic. System Card: Claude Sonnet 4.6. System card, Anthropic, February 2026d. URL https://www.anthropic.com/claude-sonnet-4-6-system-card.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, et al. AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=M7TNf5J26u.

Andrés M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8. URL https://www.nature.com/articles/s42256-024-00832-8.

ByteDance Seed. Seed2.1 model card: Agentic intelligence for productivity. Model card, ByteDance Seed, June 2026. URL https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2.1/Seed2_1_Model_Card.pdf.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, et al. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6s5uXNWGIh.

Hui Chen, Miao Xiong, Yujie Lu, et al. MLR-Bench: Evaluating AI agents on open-ended machine learning research. *arXiv preprint arXiv:2505.19955*, 2025a. URL https://arxiv.org/abs/2505.19955.

Ziru Chen, Shijie Chen, Yuting Ning, et al. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=6z4YKr0GK6.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. arXiv preprint arXiv:2606.19348, 2026. URL https://arxiv.org/abs/2606.19348v1.

GLM-5 Team. GLM-5: from vibe coding to agentic engineering. arXiv preprint arXiv:2602.15763, 2026. URL https://arxiv.org/abs/2602.15763v2.

Google DeepMind. Gemini 3.1 Pro Model Card. Model card, Google DeepMind, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026.

Xueyu Hu, Ziyu Zhao, Shuang Wei, et al. InfiAgent-DABench: Evaluating Agents on Data Analysis Tasks. In *International Conference on Machine Learning*, pp. 19544–19572, 2024.

Kexin Huang, Serena Zhang, Hanchen Wang, et al. Biomni: A General-Purpose Biomedical AI Agent. *bioRxiv*, 2025.

Qian Huang, Jian Vora, Percy Liang, and J. Leskovec. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. In *International Conference on Machine Learning*, volume 235, pp. 20271–20309, 2024. URL https://proceedings.mlr.press/v235/huang24y.html.

Sayash Kapoor and Arvind Narayanan. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv.org*, abs/2207.07048, 2022.

Sayash Kapoor, Emily F. Cantrell, Kenny Peng, et al. REFORMS: Reporting Standards for Machine Learning Based Science. *arXiv.org*, abs/2308.07832, 2023.

Kimi Team. Kimi K3: Open frontier intelligence. arXiv preprint arXiv:2607.24653, 2026. URL https://arxiv.org/abs/2607.24653v2.

Thomas Kwa, Ben West, Joel Becker, et al. Measuring AI Ability to Complete Long Software Tasks. In *Advances in Neural Information Processing Systems 38*, pp. 102382–102435. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025.

Dong Li, Si-Xuan Mi, Zi-Hao Ye, et al. Autonomous Chemical Mechanistic Discovery through Agentic Reasoning and Validation. 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv*, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vyflgpwfJW.

G. Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI Assistants. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023.

H. Nusrat and Omar Nusrat. When AI Does Science: Evaluating the Autonomous AI Scientist KOSMOS in Radiation Biology. *arXiv.org*, abs/2511.13825, 2025.

OpenAI. GPT-5.4 Thinking System Card. System card, OpenAI, March 2026a. URL https://deploymentsafety.openai.com/gpt-5-4-thinking.

OpenAI. GPT-5.6 System Card. System card, OpenAI, July 2026b. URL https://deploymentsafety.openai.com/gpt-5-6.

OpenAI. Using GPT-6. Official OpenAI API documentation, 2026c. URL https://developers.openai.com/api/docs/guides/latest-model/gpt-6-astra. Accessed September 24, 2026.

Qwen Team. Qwen3.5-397B-A17B. Official Hugging Face model card, February 2026a. URL https://huggingface.co/Qwen/Qwen3.5-397B-A17B. Accessed September 24, 2026.

Qwen Team. Qwen3.8. Official QwenLM GitHub repository and model-series documentation, August 2026b. URL https://github.com/QwenLM/Qwen3.8. Accessed September 24, 2026.

Zachary S Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark. *Transactions on Machine Learning Research*, 2025.

R. Silberzahn, E. L. Uhlmann, D. P. Martin, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 2018.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A Fine-grained and Challenging Benchmark for Process-Level Reward Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346. Association for Computational Linguistics, 2025.

Giulio Starace, Oliver Jaffe, Dane Sherburn, et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025. URL https://arxiv.org/abs/2504.01848.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, W. Tang, Alejandro Cuadron, Chenguang Wang, Raluca A. Popa, and Ion Stoica. JudgeBench: A Benchmark for Evaluating LLM-based Judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G0dksFayVq.

Yuru Wang, Lejun Cheng, Yuxin Zuo, et al. NatureBench: Can coding agents match the published SOTA of Nature-Family papers? *arXiv preprint arXiv:2606.24530*, 2026. URL https://arxiv.org/abs/2606.24530.

Yunze Wu, Dayuan Fu, Weiye Si, et al. InnovatorBench: Evaluating Agents' Ability to Conduct Innovative AI Research. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=w8rZ2Jd6Jo.

Yutaro Yamada, R. Lange, Cong Lu, Shengran Hu, Chris Lu, J. Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv.org*, abs/2504.08066, 2025.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *Advances in Neural Information Processing Systems 37*, pp. 50528–50652. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Z.ai. GLM-5.3. Official model card, Hugging Face, 2026. URL https://huggingface.co/zai-org/GLM-5.3. Accessed September 24, 2026.

Dan Zhang, Sining Zhoubian, Min Cai, Fengzu Li, Lekang Yang, Wei Wang, Tianjiao Dong, Ziniu Hu, Jie Tang, and Yisong Yue. DataSciBench: An LLM Agent Benchmark for Data Science. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3685–3728. Association for Computational Linguistics, 2026.

Chujie Zheng, Zhen-Ru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bo-Wen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying Process Errors in Mathematical Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.50/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*, pp. 46595–46623. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023.