Science or Slop?: BENCHMARKING AND MITIGATING SCIENTIFIC SLOP IN AI-GENERATED PAPERS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated content, often called AI slop, is increasingly common everywhere, particularly in academia. Slop in AI-generated scientific papers, however, has more complex patterns that cannot be easily detected by existing token-based AI detectors. Each part of such a paper looks plausible while the scientific reasoning that connects the parts breaks down, which can mislead how readers assess the work. We benchmark these failures as **scientific slop** through six measures across *Structure*, *Argument*, and *Artifacts*. We construct **SCISLOPBENCH** with 390 AI-generated papers, mostly in computer science but spanning the life, social, and natural sciences, each paired with a human-written paper matched by research problem and contribution type. Our measures identify the AI paper in each pair with 85.9% accuracy, compared with 68.7% for Binoculars. Higher scientific slop accompanies lower ICLR ratings and distinguishes rejected from accepted papers above chance in every year from 2017 to 2025. Reducing these patterns, however, is not as simple as directly optimizing the measures. We therefore propose **SCISLOPHARNESS**, a harness-level framework that guides a fixed LLM to revise slop only where the experiment records support the change. While standard revisions leave residual slop and direct slop-aware prompting triggers reward hacking, SCISLOPHARNESS reduces the remaining AI–human gap by 63% over the strongest revision baseline without requiring human reference targets. Overall, we demonstrate that AI-generated scientific papers leave fundamental traces in their global reasoning, and that responsible mitigation demands strict evidentiary grounding rather than mere prose refinement.[1]

1 INTRODUCTION

As LLM-generated content fills news feeds, books, and social media, the term *AI slop* has come to name this low-cost, high-volume output (Thorp, 2026). AI output that looks finished but lacks substance shifts the verification burden to the recipient, forcing them to review and revise the content, and raises concerns of misuse in journalism, education, workplaces, and academia (Wang et al., 2024). AI slop is increasingly common in academia, in both papers and peer reviews (Liang et al., 2025; Kobak et al., 2025; Liang et al., 2024). Paper submission platforms such as arXiv have revised their policies to restrict survey papers (arXiv, 2025), and conference organizers have begun screening submissions using AI detectors. Yet, a scientific paper is more than its *prose*.

These concerns have made detecting AI-generated content increasingly important. AI agents are now used to draft and refine scientific papers, from researcher-guided workflows (Schmidgall et al., 2025) to systems such as AI Scientist (Lu et al., 2024; Yamada et al., 2025) and FARS (Tang et al., 2026). Such assistance can reduce writing time while also increasing the need for verification, as generated scientific text can appear credible even when the underlying data are fabricated (Kacena et al., 2024; Gao et al., 2023). Existing detection methods address these concerns by identifying AI-generated content through learned textual features or token probabilities (Emi & Spero, 2024; Mitchell et al., 2023; Hans et al., 2024; Ma et al., 2026). For scientific papers, however, token-level probabilities alone are insufficient.

Challenges. We define **scientific slop** as recurring breakdowns in how scientific reasoning connects across a paper (Figure 1). Each part of a paper can look plausible in isolation, so such breakdowns

[1] Our code and datasets are available at https://anonymous.4open.science/r/ScientificSlop-01EE

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.

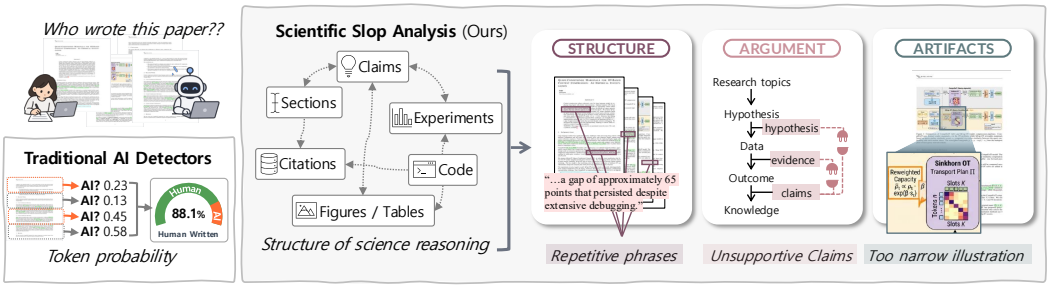


Figure 1: Scientific slop captures failures beyond token-level AI detection. While traditional detectors assess local text, scientific slop analysis examines how scientific reasoning connects, revealing paper-level failures that plausible prose can conceal.

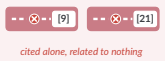
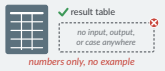
are invisible to token-level detectors and can only be identified or repaired at the level of the whole paper. To benchmark and mitigate scientific slop, this paper addresses three challenges. **First, token-level metrics fail to capture how scientific reasoning connects across a paper.** Sections, claims, citations, evidence, and artifacts can each appear plausible while the relationships among them break down. We repeatedly observe such failures in end-to-end AI-generated papers, and ICLR reviewers already penalize them even in human-written submissions. **Second, detection of these patterns remains unmeasured.** Existing test sets label only the text, so the extent to which detectors, including LLMs that read the entire paper, identify these patterns has never been measured. **Third, these patterns are difficult to mitigate reliably.** Whereas token-level signals can be removed by paraphrasing (Krishna et al., 2023), repairing these patterns requires restoring the missing relations without changing the underlying science.

To systematically study scientific slop in AI-generated papers, we construct **SciSlopBench**, the first benchmark to evaluate detection in full scientific papers rather than standalone prose. It pairs 390 AI-generated papers from FARS and Agents4Science 2025 (Bianchi et al., 2026) with human-written papers on the same research problem and of the same contribution type, mostly in computer science but also spanning the life, social, and natural sciences. Every human-written counterpart was accepted at a top-tier venue, so a system that distinguishes the two must look beyond topic, genre, and prose quality. To mitigate the identified scientific slop, we introduce **SciSlopHarness**, which translates slop diagnoses into iterative revision guidance for an LLM. In each round, it checks edits against the original manuscript and experiment records, then uses review feedback to repair remaining gaps in scientific reasoning.

Contributions. Our results formalize scientific slop as paper-level breakdowns in scientific reasoning and develop measures to quantify it. These measures reduce detection error by 55% compared with Binoculars (Hans et al., 2024), the strongest token-level baseline. Moreover, measures of scientific slop align with human evaluations: AI-likeness drops from 26% to 12% as review scores rise from 2 to 9. These measures also distinguish rejected from accepted papers in every ICLR year from 2017 to 2025. For mitigation, general revision leaves slop unresolved, while direct slop-aware revision lowers the scores without fixing the reasoning. SciSlopHarness, which repairs scientific reasoning against grounding evidence, instead moves all six measures toward human levels without using human targets. These results show that AI-generated papers leave measurable traces in how scientific reasoning connects across a paper, and that grounding revision in scientific records enables these traces to guide scientific repair. Our contributions are:

- We define and measure scientific slop through six types of breakdown in scientific reasoning across STRUCTURE, ARGUMENT, and ARTIFACTS. Individual measures achieve a pairwise accuracy of up to 0.905 in distinguishing AI-generated from human-written papers.
- We construct SciSlopBench to evaluate AI detection based on scientific reasoning across a whole paper rather than standalone prose. The strongest token-level detector, Binoculars, achieves a pairwise accuracy of only 0.687, while full-text AI reviewers achieve at most 0.685.
- We introduce SciSlopHarness, which verifies repairs against scientific records rather than slop scores. It reduces the remaining AI–human gap by 63% over the strongest revision baseline, while general revision leaves scientific slop and direct slop-aware revision overcorrects.

Table 1: Measurement specification of SciSlop patterns. Each score is the share of units that show the pattern, written as numerator / denominator (Eq. 1). The last column states what a score of 1 means. Full rules are in Appendix A.

	Item	Illustration	Unit of analysis $U_j(p)$	Score $s_j(p)$	What $s_j = 1$ means
STRUCTURE	Cross-section references		Body sections and labeled objects	objects never referenced outside their own section / all objects	No object is ever referred to from another section.
	Macro redundancy		Sentences with at least eight tokens	sentences half or more copied as 8-grams from an earlier section / all sentences	Every sentence mostly repeats earlier sections.
ARGUMENT	Argument graph		Key claims in the Introduction	shallow claims: nothing earlier leads up to them / all key claims	The argument is flat. Every claim stands alone with no build-up behind it.
	Citation isolation		Citation sentences (Intro., Related Work)	citations not grouped, compared, or related to other works / all citations	Every citation stands alone.
ARTIFACTS	Figure exposition		Content types in method figures	content types not needed to show the method / all content types in the figure	Nothing in the figure shows the method itself.
	Evidence gap		Papers with a body result table	papers showing no concrete input, output, or case / all papers	The paper gives no concrete example at all.

2 Related Work

Detecting machine generated text. Detection methods identify language model outputs by analyzing local text spans. DetectGPT measures probability curvature (Mitchell et al., 2023), Binoculars compares model perplexity (Hans et al., 2024), NTS measures sensitivity to sampling temperature (Ma et al., 2026), and Pangram trains classifiers on generated text (Emi & Spero, 2024). Watermarking and paraphrasing studies similarly focus on surface-level linguistic signals (Kirchenbauer et al., 2023; Krishna et al., 2023). Unlike these tools, our measures read relations across a whole paper rather than local text spans.

Automated review and revision. Recent systems automate the peer review process by generating critiques or revising drafts directly. The AI Scientist evaluates its own papers with a simulated review process (Lu et al., 2024), CycleReviewer learns from real peer reviews (Weng et al., 2025), and OpenReviewer and DeepReview treat review generation as a primary task (Idahl & Ahmadi, 2025; Zhu et al., 2025). Self-Refine closes the loop by having one model critique and revise its own output (Madaan et al., 2023). However, these systems offer general feedback without isolating specific recurring patterns. We use the AI Scientist reviewer and CycleReviewer as detection baselines and a reviewer-based refinement loop as a revision baseline, and ask whether general review and revision removes the flaws we define.

Measurement under active removal. Recent work improves an LLM system by changing the information, checks, and tools around a fixed model rather than the model itself (Lee et al., 2026). SciSlopHarness follows this design, keeping the editor fixed and adding located findings before the edit and a record-grounded review after it. Mutation testing (DeMillo et al., 1978; Jia & Harman, 2010) and adversarial stylometry (Brennan et al., 2012; Potthast et al., 2016) likewise judge a property by trying to remove it, and our slop-aware baseline plays that role for scientific slop.

3 Defining and Measuring Scientific Slop

3.1 Scientific Slop: Definition and Measures

Kommers et al. (2026) identify superficial competence as a characteristic of AI slop, whose apparent quality conceals a lack of substance. In scientific papers, we examine this gap in the connections needed to organize the study, substantiate its claims, and make its methods and evidence inspectable. We define **scientific slop** as recurring failures in these connections that hinder readers from following the paper's scientific reasoning. We group these failures into STRUCTURE, ARGUMENT, and ARTIFACTS. Table 1 specifies six observable patterns and their counting rules.

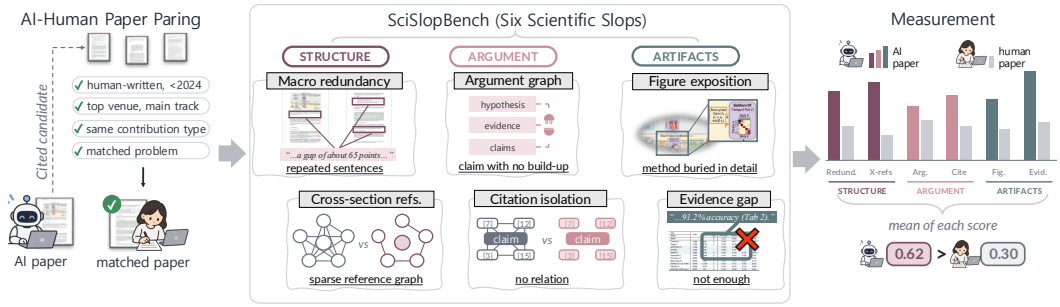


Figure 2: **SCISLOPBENCH pipeline.** We pair each AI-generated paper with a human-written paper matched by research problem and contribution type. We measure six scientific slop patterns in both papers, then compare the mean slop scores of each matched pair.

STRUCTURE concerns how sections contribute to the paper's scientific argument. *Cross-section references* targets missing connections between them. Rhetorical structure theory explains how parts of a text are connected, while endophoric references make these connections visible to readers (Mann & Thompson, 1988; Hyland, 2023). We therefore measure sections and labeled objects that no other section refers to. *Macro redundancy* captures when later sections repeat earlier material instead of developing the scientific argument. Methods, results, and conclusions serve different argumentative roles (Teufel et al., 2009), so each section should add to rather than restate the argument. We measure such reuse with cross-section n-gram overlap (Welleck et al., 2020).

ARGUMENT evaluates how a paper justifies its claims and establishes its position among prior studies. *Argument graph* targets breaks in the reasoning that connects a paper's claims to their support. Argument mining represents this reasoning through relations between claims and premises (Stab & Gurevych, 2017). We measure how often a claim appears before its supporting context. *Citation isolation* captures prior work that is cited without explaining its role in the paper's argument. Citations can show what a paper builds on, differs from, or compares against (Teufel et al., 2006). We measure citations that provide none of these relations.

ARTIFACTS examines whether readers can inspect how a method works and what its reported results represent. *Figure exposition* captures method diagrams that obscure how the method works. Diagrams support inference by showing related information together (Larkin & Simon, 1987), but unnecessary explanations can interfere with this role (Mayer & Moreno, 2003). We flag experimental details or claims about results that make the method's steps and connections harder to follow. *Evidence gap* occurs when a paper reports aggregate results but provides no concrete examples. Without these examples, readers cannot inspect individual successes and failures to better understand model behavior (Lipton & Steinhardt, 2019). We check for this gap in papers with result tables in the main text, looking for inputs, outputs, or cases throughout the paper and appendix.

For paper p and item j, let $U_j(p)$ be the units of analysis and $h_j(u)$ indicate whether unit u meets the counting criterion in Table 1. The item score is

$$s_j(p) = \frac{1}{|U_j(p)|} \sum_{u \in U_j(p)} h_j(u). \tag{1}$$

3.2 SCISLOPBENCH

Existing AI text detection benchmarks (Wang et al., 2024; Li et al., 2024; Dugan et al., 2024; Wu et al., 2024) typically evaluate standalone prose, leaving relations among a paper's sections, tables, and code untested. To study scientific slop systematically, we introduce SCISLOPBENCH, a benchmark for evaluating how well systems distinguish AI-generated papers from matched human-authored papers using whole-paper evidence. As shown in Fig. 2, SCISLOPBENCH pairs AI-generated papers with human-written papers matched by research problem and contribution type, gives each paper one slop score from the six patterns, and checks whether the human-written paper scores lower than its AI-generated counterpart.

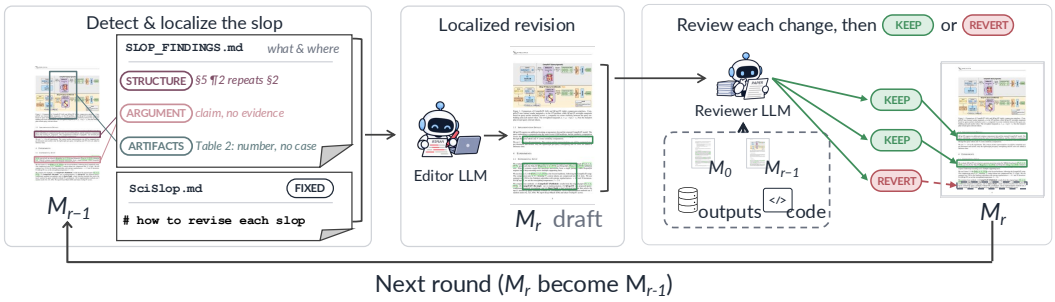


Figure 3: **SciSlopHarness pipeline.** We locate scientific slop instances in the manuscript and use them to guide localized revision. We then review each proposed change against the original manuscript, experiment outputs, and code, retain only supported changes, and repeat the process on the remaining slop.

3.2.1 Benchmark Construction

Each instance in SciSlopBench pairs an AI-generated paper with a human-authored paper matched by paper type and topic. We evaluate whether these scores rank the human-authored paper above its AI-generated counterpart. The reference labels reflect how the papers were produced.

Dataset construction. We draw AI-generated papers from FARS and Agents4Science 2025 submissions. FARS takes a research idea and produces a finished paper, running the experiments and writing the manuscript itself. For each FARS paper, we identify cited papers of the same paper type that meet our human-authorship screening criteria and have main-track acceptance at a top-tier venue. We then select the closest topical match, yielding 143 pairs matched on both the kind of contribution and the research topic. For Agents4Science, we also include papers available as PDFs, which we convert into a form suitable for slop measurement. Data collection, matching, and document-processing details are provided in Appendix B.

Benchmark statistics. SciSlopBench combines 143 matched pairs from FARS with 247 pairs from Agents4Science 2025 submissions, spanning computer and data sciences, life and health sciences, social sciences, natural sciences, and other fields. We report slop measures as normalized rates to account for differences in paper length and the number of measured units. Source-specific composition, acceptance counts, and detailed topic distributions are provided in Appendix B.

Evaluation metrics. We use PairAcc and AUROC as our main evaluation metrics. PairAcc measures how often a system ranks the human-authored paper above its AI-generated counterpart, with score directions fixed in advance and ties receiving half credit. AUROC measures how well the scores distinguish the two classes across classification thresholds. We also report the true-positive rate at a 5% false-positive rate (TPR), the share of AI papers flagged when the threshold lets through 5% of human papers. For the ICLR analysis in Section 5.2, each system's score is turned into an AI probability by a logistic fit on the FARS pairs (Appendix E).

4 SciSlopHarness: Mitigating Scientific Slop

This section describes SciSlopHarness, our method for reducing scientific slop by repairing the connections in scientific reasoning. As shown in Fig. 3, SciSlopHarness locates scientific slop, uses the detected instances to guide manuscript revision, and reviews each proposed change against the original manuscript and the paper's experiment records before retaining it. The remaining slop and review feedback then guide the next revision round.

Task. Given an original manuscript M_0 and its experiment records, the result files its code produced, SciSlopHarness revises the manuscript for up to three rounds. Following harness-level optimization (Lee et al., 2026), we keep the revision model fixed and change only the information and checks around it. At round r, the revision model receives M_{r-1} with `SciSlop.md`, which describes how each type of scientific slop should be revised. It also receives `SLOP_FINDINGS.md`,

Table 2: Paper discrimination on SCISLOPBENCH. TPR is at 5% FPR, and bold and underline mark the best two.

Method	PairAcc	AUROC	TPR
AI text detectors, prose view			
Binoculars	<u>0.687</u>	0.683	<u>0.238</u>
DetectGPT	0.638	0.623	0.146
NTS	0.626	0.607	0.123
Automated reviewers, full text			
CycleReviewer	0.685	<u>0.689</u>	0.054
AI Scientist	0.615	0.613	0.146
SciSlop (ours)	**0.859**	**0.854**	**0.272**

Table 3: Scientific slop measures by plane. TPR is at 5% FPR, a dagger marks a measure that needs a model, and the last row is the score of Table 2.

Plane	Measure	PairAcc	AUROC	TPR
STRUCTURE	Macro redundancy	0.723	0.727	<u>0.467</u>
	Cross-section refs.	**0.905**	**0.895**	**0.648**
ARGUMENT	Argument graph†	0.586	0.584	0.159
	Citation isolation	0.793	0.782	0.444
ARTIFACTS	Figure exposition†	<u>0.809</u>	<u>0.786</u>	0.364
	Evidence gap	0.764	0.765	0.000
SciSlop (ours), mean of planes		0.859	0.854	0.272

which lists the detected problems and their locations. The revision model edits the manuscript to produce a draft $\widetilde{M}_r$. A review model then compares each change with M_{r-1}, M_0, experiment outputs, and the code. Changes consistent with the reported results, existing evidence, and cited work are kept in M_r. The remaining changes are reverted to their version in M_{r-1}, so each round proceeds as $M_{r-1} \xrightarrow{\text{revision}} \widetilde{M}_r \xrightarrow{\text{review}} M_r$. Appendix C provides implementation details.

4.1 LOCALIZED REVISION

We first locate where a connection in scientific reasoning breaks down so that the revision model knows what part of the manuscript to revise. For each detected problem, the model receives its location and guidance on how to revise it. The locations come from measurements of the manuscript M_{r-1} and are stored in `SLOP_FINDINGS.md`. The revision guidance is fixed across manuscripts and stored in `SciSlop.md`. The model uses both to edit the LaTeX source and produce $\widetilde{M}_r$.

We use these findings to guide revision, not to require every scientific slop measure to decrease. The model can leave a finding unchanged when the suggested connection is not needed in the paper. For example, it adds a cross-section reference only when another section actually uses the referenced object. We therefore provide the findings and revision guidance without a target score.

4.2 REVIEWING PROPOSED REVISIONS

Localized guidance tells the revision model what to change, but it cannot verify whether the new content is supported by the research. We therefore review each proposed change before keeping it in the manuscript. For each changed paragraph in $\widetilde{M}_r$, the review model checks the original manuscript, experiment outputs, and code for the claims, numbers, citations, and examples affected by the revision. It keeps a change only when these sources support the revised content. Otherwise, the paragraph is restored to its version in M_{r-1}. The remaining changes form M_r.

The review checks that revisions preserve the paper's reported results, citations, and available evidence. A changed result is reverted if it conflicts with the experiment outputs. A new example is kept only when the same example exists in the available records. Changes that remove citations or add unnecessary references are also reverted.

We repeat revision and review for up to three rounds. We stop when no problems remain for revision or when review produces no further changes. Before returning the final manuscript, we check that reported numbers and cited works are preserved and that citations and references remain valid.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Baselines. All existing methods assign a single score to each paper. We evaluate Binoculars (Hans et al., 2024), DetectGPT (Mitchell et al., 2023) and NTS (Ma et al., 2026) as AI text detectors, and

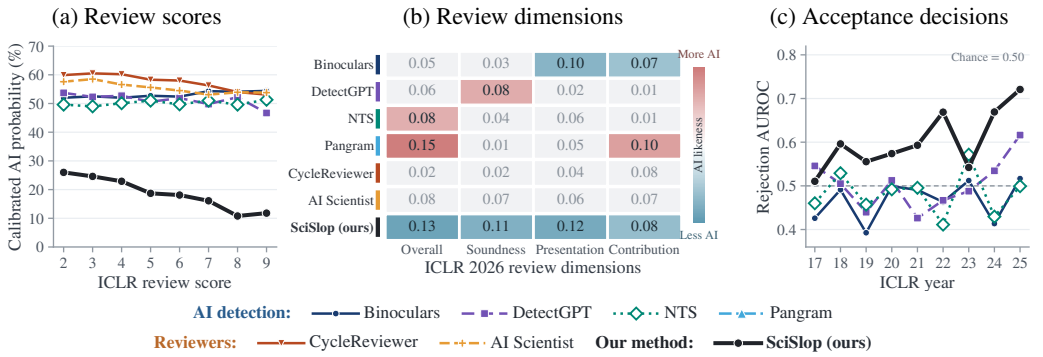


Figure 4: AI likeness against ICLR review scores (a), assessment dimensions (b), and annual acceptance decisions (c). Panel (a) adjusts probabilities for differences between years, panel (b) shows $|\rho|$ with blue and red for significant negative and positive correlations ($p < 0.05$) and gray for estimates that are not significant, and panel (c) uses rejection AUROC with 0.5 indicating chance.

use the overall ratings from the AI Scientist reviewer (Lu et al., 2024) and CycleReviewer (Weng et al., 2025) as automated review baselines. For revision, we compare four baselines that share the editor model and round protocol. Base prompting, Claude Code, and reviewer-based refinement receive only a generic improvement instruction, and we refer to these three as general revision. Slop-aware revision instead receives explicit slop definitions and detected locations, using the same editor and tool configuration as reviewer-based refinement. We compare these baselines with our SciSlopHarness, using the negative of the aggregate item score so that higher values favor human authorship, as for the other methods. Details of the revision baselines are provided in Appendix D, and the distances behind the reported gap reduction in Appendix C.5.

Implementation details. Text detectors read the prose view. Other methods read the paper source, with code access provided where needed. Binoculars uses the Falcon-7B (Almazrouei et al., 2023) base and instruct models, while DetectGPT uses t5-3b (Raffel et al., 2020) to generate 100 perturbations. For inputs exceeding Falcon's 2048-token context limit, we score windows without overlap and average their scores. The AI Scientist reviewer uses Qwen2.5-32B-Instruct (Yang et al., 2024), and CycleReviewer uses its released 8B checkpoint.

5.2 Does Scientific Slop Mark Provenance, Quality, or Both?

We test whether scientific slop captures overlooked weaknesses already reflected in lower human review scores. We first assess how well slop measures identify AI-generated papers, then examine whether the same patterns matter to human peer review.

Scientific slop identifies AI-generated papers by what detectors and reviewers never read. Tables 2 and 3 show that scientific slop identifies AI-generated papers more accurately than the listed detectors and automated reviewers that read the full paper. The best of each reaches 0.687 and 0.685, while scientific slop reaches 0.859, cutting the error by more than half. *Cross-section references* alone reaches 0.905 without a model and, at 5% FPR, flags 65% of AI-generated papers against 24% for Binoculars. These results show that measuring patterns across a paper provides an effective signal for AI detection on this corpus. By defining these patterns as specific defects, scientific slop further connects detection to concrete targets for revision.

Scientific slop reads human-written papers the way their reviewers do. Figure 4(a) compares each system's AI probability with human review ratings. Scientific slop shows a clearer **relationship with review scores** than the evaluated detectors and automated reviewers. Its AI probability declines more sharply as ratings rise, indicating that the score also captures differences in reviewer-assessed quality. Figure 4(b) examines individual review dimensions. Only scientific slop consistently assigns higher scores to papers rated lower across overall rating, soundness, presentation, and contribution, with statistical support in all four dimensions. This links scientific slop to **soundness and contribution** as well as presentation. Figure 4(c) compares the separation of rejected and accepted papers

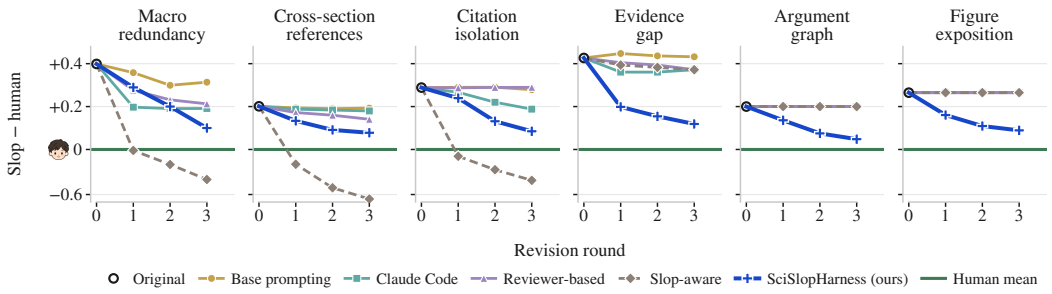


Figure 5: Scientific slop across revision rounds for SCISLOPHARNESS and baselines. Values closer to zero (green) are closer to matched human means, with round 0 denoting original manuscripts and the scale below zero compressed.

across years. Scientific slop exceeds chance in all nine years, whereas detectors fall below chance in most comparisons. The relationship thus extends to **acceptance decisions** across changing review scales. Together, these findings suggest that the weaknesses we define as scientific slop are already reflected in lower human evaluations. Scientific slop makes these recurring weaknesses explicit and measurable. Measurement details are provided in Appendix E.

5.3 DOES SCISLOPHARNESS REMOVE THE SLOP THAT GENERAL REVISION LEAVES?

We evaluate whether SCISLOPHARNESS removes scientific slop that survives general revision without receiving human anchors as revision targets. We run SCISLOPHARNESS and the four revision baselines on every AI-generated paper in SCISLOPBENCH, revising each paper for three rounds under the shared protocol of Appendix D.1. We first compare it with general revision, then examine whether repairs guided by connections in scientific reasoning approach the reference levels observed in human papers.

SCISLOPHARNESS reduces the AI–human gap that general revision leaves and slop-aware revision overcorrects. Figure 5 shows that SCISLOPHARNESS ends closest to the human mean on all six measures, despite never receiving these means as revision targets. For *Macro redundancy*, *Cross-section references*, and *Citation isolation*, general revision leaves substantial slop, whereas slop-aware revision pushes scores past the human means and increasingly away from them, suggesting overcorrection. SCISLOPHARNESS reduces the defects left by general revision while remaining closer to human levels. Its gains also extend to *Evidence gap*, *Argument graph*, and *Figure exposition*, where the baselines make little or no progress. The harness evaluates each defect in the context of the paper's scientific reasoning. The skill guides edits that connect claims with supporting evidence where needed, while review checks whether each edit contributes to the argument and preserves existing information. Together, they address defects left by feedback alone while limiting overcorrection. Summed over the six measures, the remaining gap to the human means is 63% smaller than for Claude Code, the closest baseline (Appendix C.5). The approach toward human levels thus supports scientific reasoning as a criterion for deciding what to repair and what to preserve, without specifying a target score.

5.4 IS THE SLOP REMOVED, OR ONLY THE SCORE?

We test whether reductions in scientific slop reflect valid repairs rather than lower scores alone. We compare the revisions of SCISLOPHARNESS with slop-aware revision to see what produces these reductions, and ablate its components.

SCISLOPHARNESS removes the slop where slop-aware revision removes only the score. Table 4 shows how slop-aware revision and SCISLOPHARNESS repair the same detected problems differently. For Macro redundancy, slop-aware revision reaches a zero score by removing the repeated sentence together with four citations, losing the attribution to prior work. In contrast, SCISLOPHARNESS removes only the repetition and preserves all four citations, reducing redundancy without losing the information needed for the argument. For Cross-section references, slop-aware revision

Macro redundancy		Shaded sentence repeats the Abstract. × Slop-aware drops the citations. ✓ SCISLOPHARNESS keeps them.
Original		RLVR has emerged as a powerful paradigm… from early work on CodeRL [1] and RLTF [2] to recent breakthroughs like DeepSeek-R1 [3] and DeepSeekMath [4]…
Slop-aware	×	… Early work in Section [Related Work] demonstrates… from pioneering applications of unit test signals to recent breakthroughs like DeepSeek-R1 and DeepSeekMath…
SCISLOPHARNESS	✓	In code generation, unit tests provide a natural verifier… from early work on CodeRL [1] and RLTF [2] to recent breakthroughs like DeepSeek-R1 [3] and DeepSeekMath [4]…
Cross-section references		Table [tau] is unreferenced. × Slop-aware scatters references. ✓ SCISLOPHARNESS adds one where shaded.
Original		We presented Delta-Prefill Switching (DPS)… DPS achieves 21–22% speedup over greedy decoding… The method is robust to threshold selection and requires no model modifications.
Slop-aware	×	As positioning against prior work (Section [Related Work]) shows… Delta-Prefill Switching (Section [Method]) addresses a fundamental tension… Our experiments (Section [Experiments]) demonstrate… Sequential performance (Table [sequential]) shows DPS achieves 21–22% speedup… concurrent evaluation (Table [concurrent] and Figure [concurrent]) reveals… Threshold sensitivity analysis (Table [tau] and Figure [tau]) confirms robustness across a wide range of parameter values.
SCISLOPHARNESS	✓	We presented Delta-Prefill Switching (DPS)… DPS achieves 21–22% speedup over greedy decoding… The method is robust to threshold selection (Table [tau] demonstrates that all $\tau \geq 32$ outperform the greedy baseline) and requires no model modifications.

Table 4: The same passage before revision, after slop-aware revision, and after SCISLOPHARNESS, marking the located slop, what a revision lost or left, and what it kept or repaired. Brackets replace citation and reference commands.

approaches zero by adding references to three sections, three tables, and two figures throughout the conclusion. These references improve the score, but most only tell the reader where to look without connecting the evidence to a claim. In contrast, SCISLOPHARNESS adds one reference where the threshold table supports the robustness claim, and states what the table shows, that every $\tau \geq 32$ beats the greedy baseline. SCISLOPHARNESS reduces the same slop by preserving necessary information and repairing the missing connections between claims and evidence. We observe the same distinction for the remaining measures in Appendix F.

The full SCISLOPHARNESS is needed for reliable slop reduction. Figure 6 shows that removing any part of the harness introduces a different failure. Definitions and locations reduce the distance to 0.11, but without review this comes with guard breaks in 23% of rounds. Review alone leaves most slop unrepaired at 0.30 and still breaks a guard in 17% of rounds. Definitions with review are safer, but without located instances the distance remains 0.25. In contrast, the full SCISLOPHARNESS reduces the distance to 0.17 with no guard breaks. The ablations show why the components must work together: the harness must identify where to repair, guide the repair, and verify that the resulting edit preserves the scientific record.

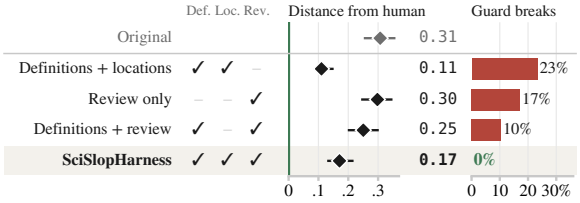


Figure 6: Ablation of SCISLOPHARNESS.

6 DISCUSSION

Conclusion. We introduce **scientific slop** as measurable breakdowns in the reasoning that connects a paper's parts. Its measures distinguish AI-generated from human papers and are associated with lower scores across all four ICLR review dimensions. SCISLOPHARNESS uses the identified defects to guide revisions checked against the paper's own experiment outputs and code. It brings all six measures closest to human levels among the evaluated revision methods, without receiving human reference targets. Scientific slop connects detection to weaknesses reflected in human reviews and to repairs supported by scientific records.

Limitations and future directions. The six scientific slops offer an initial view of how connections within a paper reveal defects and guide repair. Future work can extend this view to how a study's question, methods, and evidence support its conclusions as a whole. Our evaluation focuses on FARS and Agents4Science 2025, with most papers drawn from computer science. Extending the paired evaluation to new generators and disciplines can test how broadly this view applies.

AI USE STATEMENT

LLMs are used in this work both as the subject of study and as components of our experiments. SciSlopBench includes third-party AI-generated papers, our measures use Qwen2.5 models, and SciSlopHarness and its baselines use Claude models as editors and reviewers. We also used Claude (including Claude Code) and ChatGPT during implementation and manuscript preparation. They assisted with writing and debugging code for the measurement pipeline, harness, baselines, and analysis scripts, as well as editing the manuscript for grammar, clarity, and concision. In some cases, they were used to produce initial text that was subsequently edited by the authors. The authors reviewed and tested all assisted code and reviewed the final manuscript. All reported results were obtained by running the released code.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to support scientific scrutiny by making failures in reasoning identifiable and open to inspection. Scientific slop measures give authors and reviewers concrete points to examine in a paper, including weaknesses that fluent prose can obscure. These findings concern the paper's reasoning and do not establish an author's intent or misconduct.

Correcting a reasoning failure remains valuable even when it weakens a signal of AI authorship. The criterion for responsible repair is whether the study's records support the change. SCISLOPHARNESS applies this criterion by checking revisions against the manuscript and experiment records and recording preservation checks for citations and reported numbers. When no concrete example exists in the records, the revision acknowledges that absence. This makes the basis for a revision available for inspection alongside its effect on the score.

Our ICLR analysis uses public manuscripts and review information and reports aggregate relationships between measured slop and review outcomes. The benchmark's data sources and processing procedures are documented to support scrutiny of its construction.

REFERENCES

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867.

Anthropic. Claude code. https://claude.com/claude-code, 2025.

arXiv. Attention authors: Updated practice for review articles and position papers in arXiv CS category. https://blog.arxiv.org/2025/10/31/attention-authors-updated-practice-for-review-articles-and-position-papers-in-arxiv-cs-category/, October 2025. Accessed 2026-09-21.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

Federico Bianchi, Owen Queen, Nitya Thakkar, Eric Sun, and James Zou. Exploring the use of AI authors and reviewers at Agents4Science. volume 44, pp. 11–14, 2026.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22, 2012.

Richard A DeMillo, Richard J Lipton, and Frederick G Sayward. Hints on test data selection: Help for the practicing programmer. *Computer*, 11(4):34–41, 1978.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, 2024.

Bradley Emi and Max Spero. Technical report on the pangram ai-generated text classifier, 2024.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, 6(1):75, 2023.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Ken Hyland. *Metadiscourse*. Routledge, 2023.

Maximilian Idahl and Zahra Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, 2025.

Yue Jia and Mark Harman. An analysis and survey of the development of mutation testing. *IEEE transactions on software engineering*, 37(5):649–678, 2010.

Melissa A Kacena, Lilian I Plotkin, and Jill C Fehrenbacher. The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports*, 22(1):115–121, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International conference on machine learning*, pp. 17061–17084. PMLR, 2023.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27): eadt3813, 2025.

Cody Kommers, Eamon Duede, Julia Gordon, Ari Holtzman, Tess McNulty, Spencer Stewart, Lindsay Thomas, Richard Jean So, and Hoyt Long. Why slop matters. *ACM AI Lett.*, 1(1), March 2026. doi: 10.1145/3786777. URL https://doi.org/10.1145/3786777.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27469–27500, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

Jill H. Larkin and Herbert A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cogn. Sci.*, 11(1):65–100, 1987. doi: 10.1111/J.1551-6708.1987.TB00863.X. URL https://doi.org/10.1111/j.1551-6708.1987.tb00863.x.

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-harness: End-to-end optimization of model harnesses. *arXiv preprint arXiv:2603.28052*, 2026.

Yafu Li, Qintong Li, Leyang Cui, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, Yue Zhang, et al. Mage: Machine-generated text detection in the wild. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (Volume 1: Long Papers)*, pp. 36–53, 2024.

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. 2024.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, 9(12):2599–2609, 2025.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship, 2019. URL https://doi.org/10.1145/3317287.3328534.

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

Shixuan Ma, Jiahao Li, Zhendong Mao, and Quan Wang. Zero-shot detection of LLM-generated text using temperature sensitivity. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37664–37679, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023.

William Mann and Sandra Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8:243–281, 01 1988. doi: 10.1515/text.1.1988.8.3.243.

Richard E. Mayer and Roxana Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1):43–52, 2003. doi: 10.1207/S15326985EP3801_6.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pp. 24950–24962. PMLR, 2023.

Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*, 2023.

Martin Potthast, Matthias Hagen, and Benno Stein. Author obfuscation: Attacking the state of the art in authorship verification. pp. 716–749, 2016.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

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. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, 2025.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Comput. Linguistics*, 43(3):619–659, 2017. doi: 10.1162/COLI_A_00295. URL https://doi.org/10.1162/COLI_a_00295.

Qiong Tang, Tianxiang Sun, Xiangkun Hu, Xiangyang Liu, Yiran Chen, Yunfan Shao, Bobo Li, Changze Lv, Cheng Xu, Chengsong Huang, et al. Fars: A fully automated research system deployed at scale. *arXiv preprint arXiv:2606.31651*, 2026.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In Dan Jurafsky and Éric Gaussier (eds.), *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pp. 103–110. ACL, 2006. URL https://aclanthology.org/W06-1613/.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 1493–1502, 2009.

H Holden Thorp. Resisting AI slop. *Science*, 391(6780):5–5, 2026.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1369–1407, 2024.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *International Conference on Learning Representations*, volume 2025, pp. 3669–3709, 2025.

Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S Chao. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems*, 37:100369–100401, 2024.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2024. URL https://arxiv.org/abs/2412.15115.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. CHEAT: A large-scale dataset for detecting ChatGPT-written abstracts. *arXiv preprint arXiv:2304.12008*, 2023.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, 2025.

A Measurement Specification of the Slop Measures

This appendix provides the detailed measurement procedure for each measure in Table 1.

Document views. We expand all `\input` files and read each paper as a single LaTeX document. We define the paper body as the text before the appendix, acknowledgements, or bibliography, whichever comes first, and apply the same rule to both papers in each pair. We construct two views of this body. The *structure view*, used by the six scientific slop measures, preserves sections, labels, references, citations, and floats. The *prose view*, used by the text detectors, removes floats and captions and replaces mathematics, citations, and references with sentinel tokens. For human papers, we construct these views directly from the LaTeX e-print. Among the 247 Agents4Science submissions, 18 provide manuscript source. For the remaining 229, we reconstruct the same LaTeX structure from the PDF, recovering headings, float labels, citations, and references.

Macro redundancy. A sentence is recycled when more than half of its tokens sit inside 8-grams that an earlier section already used. Over the prose view of the abstract and the body sections, for a sentence u of at least eight tokens, with $C(u)$ the tokens of u that lie in an 8-gram first seen in an earlier sentence of a different section,

$$h(u) = \mathbf{1}[\,|C(u)|/|u| \geq 0.5\,]\,. \tag{2}$$

Cross-section references. Every body section, and every figure, table, equation, algorithm, or theorem labeled inside one, is an object the paper declares. $R(o)$ collects the sections other than its home that point to o through a reference command, a printed section number, or a named reference, and

$$h(o) = \mathbf{1}[\,R(o) = \emptyset\,]\,. \tag{3}$$

Pointing at a table from the section that holds it is neither rewarded nor punished, since only other sections count. Pointers into the appendix, and a roadmap in the first section, the longest run of forward pointers to later sections, are recorded apart because neither is a trace of one section having read another.

Argument graph. Does the context that supports a claim come before the claim? A language model labels each sentence of the introduction, by a forced choice over the numbered introduction with the majority of three greedy runs, and the key claims are the sentences labeled superiority, prior limitation, or design choice. The supporting context of s_i is the sentence that raises its likelihood most under a second model that is never prompted,

$$\begin{aligned} \mathrm{pmi}(j \to i) &= \frac{1}{|s_i| - 1} \sum_{t=2}^{|s_i|} \Big[\, \log P\big(s_i^t \mid s_j,\, s_i^{<t}\big) - \log P\big(s_i^t \mid s_i^{<t}\big) \,\Big], \\ \pi(i) &= \arg\max_{j \neq i} \mathrm{pmi}(j \to i), \end{aligned} \tag{4}$$

and the claim is marked when that context comes after it,

$$h(i) = \mathbf{1}[\,\pi(i) > i\,]\,. \tag{5}$$

Citation isolation. A citing sentence weaves when it cites two works, names a second work outside a citation command, or links two works by a cue from a frozen list, and it is isolated otherwise,

$$h(u) = \mathbf{1}[\,u \text{ does not weave}\,]\,. \tag{6}$$

The score runs over the citing sentences of the introduction and the related work. Relating a work only to the present paper is not weaving, and fewer than eight citing sentences flags the score weak.

Figure exposition. The method figure is read as text. One vision model transcribes it, with the same prompt on both sides, and six kinds of exposition are checked on the transcript, a notation key, a comparison arm, experimental content, an evaluative mark, a thesis box, and enumerated stages, with $h(k) = 1$ when the figure carries kind k. Five kinds are frozen patterns. Experimental content, whether a number is a setting of the run or a constant of the method, is decided by one reader, on 33 clear and 25 borderline figures of the FARS half and 13 and 9 of the Agents4Science half. On the FARS half the method figure is the pipeline's overview image, and on the Agents4Science half both sides are located by the same caption gate.

Evidence gap. Here the unit is the paper. It is applicable when its body holds a result table with at least two data rows and four numeric cells, and $h(p) = 1$ when no exhibit appears anywhere in it, appendix included. An exhibit is a display environment carrying consumed or produced material, a caption announcing an example, a case, or a failure, or a long quotation outside the related work, and one exhibit anywhere closes the gap.

Aggregate score. The SciSlop score of a paper averages the item scores of Eq. 1 within each plane and then averages the planes. With $\mathcal{J}_k(p)$ the items of plane k that apply to paper p and $\mathcal{K}(p)$ the planes with at least one such item,

$$S(p) = \frac{1}{|\mathcal{K}(p)|} \sum_{k \in \mathcal{K}(p)} \frac{1}{|\mathcal{J}_k(p)|} \sum_{j \in \mathcal{J}_k(p)} s_j(p). \tag{7}$$

This is the SciSlop row of Tables 2 and 3; the negative of $S(p)$ is used wherever a higher score should favor human authorship.

Models and serving. *Argument graph* labels sentences with Qwen2.5-32B-Instruct (Yang et al., 2024) and reads probabilities from Qwen2.5-7B-Instruct, and *Figure exposition* transcribes with Qwen2.5-VL-32B-Instruct (Bai et al., 2025). Models run on vLLM (Kwon et al., 2023) in bfloat16 with greedy decoding, a label or scalar is the majority or median of three runs, every call is cached under a hash of prompt and model, and no measurement calls a paid or closed model.

B SciSlopBench Construction Details

SciSlopBench holds 390 pairs. Each pair is one AI-generated paper and one human-written paper, its anchor, of the same kind of contribution on the same or a neighboring problem, and every anchor is a paper whose human authorship and peer acceptance can be shown. The AI papers come from two sources, 143 written end to end by FARS (Tang et al., 2026) and the 247 submissions to Agents4Science 2025, and the same pairing rule is applied to both. Tables 2 and 3 pool the 390 pairs into one evaluation. Appendix B.1 gives the pairing rule, Appendix B.2 the FARS half, Appendix B.3 the Agents4Science half and how its PDFs are read, and Appendix B.4 the relation to existing detection benchmarks.

B.1 Pairing rule

A human paper can anchor an AI paper only if it passes four conditions.

- H1 **Human authorship.** Its first arXiv version is dated 2024 or earlier, or it is an ICLR 2026 submission whose Pangram AI share is at most 0.05. The date stands in for a check we cannot run on older papers, and Pangram is the only check available for papers written after 2024.
- H2 **Acceptance.** It was accepted to the main track of a top-tier venue, shown by the venue record, the arXiv comment or journal reference, or OpenReview. Rejected submissions, workshop papers, and preprints do not pass.
- H3 **Contribution type.** It is the same kind of paper as the AI paper, one of method, framework, benchmark, dataset, analysis, or survey.
- H4 **Usable source.** Its arXiv e-print expands in document order to a body of at least 2,500 words in at least four sections, with the appendix separable from the body.

Within each contribution type, we select the human paper closest to the AI paper in research problem. A language model assigns topic similarity from the title and abstract on four tiers: 3 for the same problem and contribution type, 2 for the same problem, 1 for the same area, and 0 for papers cited only as tools, models, or datasets. We select the highest tier and break ties by experiment-section citation, availability of public review scores, a human-to-AI body-length ratio within 2.5, and finally SPECTER2 cosine similarity (Singh et al., 2023). We use cosine only as the final tie-break because candidate similarities are tightly clustered between 0.90 and 0.97.

We search two candidate pools in order. We first consider papers cited by the AI paper, since these are the works against which the paper positions its contribution. If no cited paper satisfies the pairing conditions, we search 2,690 accepted ICLR 2026 papers with an arXiv identifier and Pangram share at most 0.05, retrieve the three nearest by TF-IDF over title and abstract, and apply the same criteria. Assignment is one-to-one within each source, with contested anchors assigned to the closer match. We retain the candidate pool, similarity tier, condition outcomes, and alternative candidates for each pair so that every match can be independently checked.

B.2 FARS: collection and matching

Source. FARS contains 166 papers generated end to end from a research idea, including experiments and manuscript writing. Of these, 165 provide LaTeX source and 160 provide code. Most are method papers (123), followed by analysis (38), framework (4), and benchmark (1).

Matching. Of the 166 FARS papers, 161 cite at least one arXiv paper, yielding 1,946 candidates from 1,197 unique papers. Applying H1–H4 leaves candidates for 124 papers. H3 removes the most because many citations are benchmarks or base models rather than papers of the same contribution type. One-to-one assignment then yields 121 pairs from the cited pool, resolving 16 shared-anchor

conflicts in favor of the closer match. For the remaining 45 papers we search the widened ICLR 2026 pool, which pairs 23 more; the other 22 (13 analysis, 7 method, 2 framework) have no same-type anchor in either pool and are not paired. One of the 144 pairs is dropped because its FARS paper ships no LaTeX source, leaving the 143 pairs of the benchmark, 120 from the cited pool and 23 from the widened pool (Table 5).

Stratum	Pairs	Mean cosine	Rank 1	Rank ≤ 3	Rank ≤ 10	Median rank
All pairs	**143**	**0.927**	37	71	101	4
Cited pool	120	0.928	33	61	85	3
Widened pool	23	0.919	4	10	16	5
Cross-pair reference	20,306	0.881				

Table 5: Retrieval audit of the assignment with SPECTER2 similarity between the AI paper and the human anchor. Rank is the position of the assigned anchor among all 143 anchors for that AI paper, and the reference row gives the mean cosine over every non-assigned pair.

B.3 Agents4Science: collection, matching, and document processing

Collection. Agents4Science 2025 provides 247 primarily AI-authored submissions, and we include all of them. Most are in computer and data sciences (178), particularly AI and machine learning (154). By contribution type, 171 are method papers, 46 analysis, 19 framework, 6 survey, and 5 benchmark. Acceptance decisions, reviewer scores, and reported AI involvement are retained as metadata but are not used for pairing. Tables 6 and 7 report the full topic distribution.

Matching. Every one of the 247 submissions is paired with a human anchor that passes H1, H2, and H4. The full pairing rule, which also requires the same contribution type, pairs 185 of them: 66 anchors cited by the submission and 119 from the widened ICLR 2026 pool at tier 2 or higher. The remaining 62 submissions cover topics poorly represented in ICLR, including physics, astronomy, biology, medicine, finance, and human-computer interaction, so for these we keep H1, H2, and H4 fixed and take the closest admissible anchor in a recorded order. Of the 62, 19 share the research problem, 21 share the contribution type and the area, 16 share the area, and 6 are matched to the nearest TF-IDF candidate. The last 22 are thus still paired with an accepted, human-written paper, at the area level or by nearest neighbor, rather than left without a counterpart. Overall, 206 of the 247 anchors match the contribution type and 196 match the research problem. We record the match level for every pair, allowing evaluation on the strict 185 pairs, the 196 problem-matched pairs, or all 247.

Document processing. We read human anchors from their arXiv LaTeX source. Among the AI submissions, 18 have LaTeX source consistent with the submitted PDF and are read directly; the remaining 229 are reconstructed from PDF into the LaTeX elements required by our measures. The reconstruction extracts section headings, figure and table captions, citations, cross-references, and result-table cells using PyMuPDF. It does not infer content absent from the PDF. On the 18 submissions available in both formats, reconstruction recovers a median of 20 body headings compared with 21 in the source. We record the processing route for every paper.

B.4 Comparison with existing benchmarks

SciSlopBench is designed to evaluate scientific failures that passage-level detection benchmarks cannot represent (Table 8). Existing benchmarks primarily treat passages or abstracts as independent text instances. In contrast, SciSlopBench preserves complete papers and pairs each AI paper with a human paper addressing the same research problem. This makes relations across sections, claims, citations, figures, tables, and evidence directly measurable rather than reducing detection to local writing patterns. The FARS half further includes the code that produced the reported results, enabling checks between manuscript claims and experimental records. While existing benchmarks provide broader coverage of genres and adversarial rewrites, SciSlopBench provides the paper-level structure and matched scientific context required to evaluate scientific slop.

Primary topic	Total	Acc.	Rej.
Computer & Data Sciences	178	39	139
Life & Health Sciences	17	2	15
Social Sciences	14	5	9
Natural Sciences	13	1	12
Interdisciplinary	11	0	11
Engineering & Technology	6	1	5
Mathematics & Statistics	4	0	4
Humanities & Arts	2	0	2
Law, Policy & Business	1	0	1
Not specified	1	0	1
Total	**247**	**48**	**199**

Table 6: Primary topics and acceptance status of the 247 Agents4Science 2025 papers listed in the public submission table (Acc./Rej.: accepted/rejected). One rejected submission has no domain annotation.

Primary / secondary topic	Total	Acc.	Rej.
Computer & Data Sciences			
AI & Machine Learning	154	32	122
Human-Computer Interaction	19	6	13
Data Science	3	1	2
Software & Systems	2	0	2
Life & Health Sciences			
Medicine & Health	10	0	10
Biology	4	2	2
Agriculture & Ecology	2	0	2
Neuroscience	1	0	1
Social Sciences			
Economics	5	3	2
Psychology	5	1	4
Communication & Media	2	0	2
Operations Research	1	1	0
Education	1	0	1
Natural Sciences			
Physics	9	0	9
Astronomy	4	1	3
Interdisciplinary			
Cognitive Science	5	0	5
Computational Social Science	4	0	4
Science & Technology Studies	1	0	1
Biomedical Engineering	1	0	1
Engineering & Technology			
Materials Science	4	1	3
Biomedical Engineering	1	0	1
Electrical & Computer Engineering	1	0	1
Mathematics & Statistics			
Pure & Applied Mathematics	3	0	3
Operations Research	1	0	1
Humanities & Arts			
Philosophy	1	0	1
History	1	0	1
Law, Policy & Business			
Public Policy	1	0	1
Not specified	1	0	1
Total	**247**	**48**	**199**

Table 7: Secondary topics within each primary category for the 247 Agents4Science 2025 papers listed in the public submission table. Rejected counts are the difference between total and accepted counts. Human-Computer Interaction combines two punctuation variants. Biomedical Engineering and Operations Research each appear under two primary categories.

Benchmark ↓ Property →	Scientific text	Whole paper	Tables and figures	Released code	Executed research	Matched human pair	Multiple genres	Adversarial variants
RAID (Dugan et al., 2024)	✓	–	–	–	–	✓	✓	✓
DetectRL (Wu et al., 2024)	✓	–	–	–	–	◯	✓	✓
M4 (Wang et al., 2024)	✓	–	–	–	–	✓	✓	–
MAGE (Li et al., 2024)	✓	–	–	–	–	◯	✓	–
CHEAT (Yu et al., 2023)	✓	–	–	–	–	✓	–	◯
IDMGSP (Mosca et al., 2023)	✓	◯	–	–	–	✓	–	◯
SciSlopBench (ours)	✓	✓	✓	✓	✓	✓	–	△

Table 8: Comparison with existing machine-generated text benchmarks. (✓) the benchmark instances carry the property, (◯) partly, such as a human text from the same source rather than a matched counterpart, (△) studied in the paper but not in the benchmark instances, (–) absent.

C SciSlopHarness Implementation Details

This appendix gives the implementation of SciSlopHarness used for Figure 5. Table 9 lists the configuration, and the following subsections describe each stage in the form of the harness summaries of Lee et al. (2026). Figure 5 reports every AI-generated paper in SciSlopBench, each revised for three rounds, and the baselines of Appendix D are run on the same papers.

C.1 Configuration

The editor, the six definitions, and their measurers stay fixed across all conditions, and so does the reviewer. A round consists of one editor pass, one review pass, and one measurement.

Setting	Configuration used for Figure 5
Editor	`claude-haiku-4-5-20251001`
Reviewer	`claude-sonnet-5`, no tools
Retirement call	`claude-haiku-4-5-20251001`
Skill	`SciSlop_v0.5.md`, staged as `SciSlop.md`
Budget	At most three rounds
Measured patterns	All six
Figure editing	Located expository phrases erased from the raster method figure by inpainting. The edit is a reviewed change.

Table 9: Configuration of the default revision procedure.

C.2 Localized revision

Overview. The editor receives two files. `SciSlop.md` says how each pattern is repaired, and `SLOP_FINDINGS.md` says where the remaining instances are. It revises one source file per call. Table 10 lists the repair direction of each pattern.

- **Skill file.** `SciSlop.md` holds the six definitions, each with a repair direction and the format of its located instances. Its global rules are short. Do not fabricate experiments, numbers, citations, or examples. Do not add or remove cited works and do not change reported numbers. Keep the method, the experiments, and the results as they are. Make the smallest edit that removes the pattern. Treat listed locations as candidates and leave one alone when the repair would add text the argument does not need. The file has no score formula and no target.
- **Located feedback.** `SLOP_FINDINGS.md` lists the remaining instances by pattern, up to 8 per pattern with the count of the rest. From round 2 onward, the previous round's review notes are appended: the reverted changes with the reviewer's reasons, and the retired instances.
- **Editor call.** For each of `main.tex` and `sections/*.tex`, one call sends `SciSlop.md`, `SLOP_FINDINGS.md`, the other files as read-only context, and the file to edit. The editor edits only its own file. It receives no human paper, no score, and no formula.
- **Edit blocks.** The call returns edit blocks, each quoting an original passage and giving its replacement, or the word `NOCHANGE`. A block is applied only when the quoted passage occurs exactly once in the file. An edited file is accepted only if it keeps the file's structural commands and its length stays within 0.5 to 2.5 times the original.
- **Figure editing.** The method figure is a raster image, so the editor lists the expository phrases to erase and a deterministic step erases them. The controller keeps only phrases the measurer had classified as expository. A vision-language model returns a bounding box for each phrase and an inpainting model fills the box from its surroundings. An edit report records the erased phrases, the boxes, and the pixels changed outside them.
- **Records.** Before each round, the controller copies the project's own experiment records from `exp/EXPERIMENT_RESULTS` into `materials/`. Once per paper, a specimen search checks whether these records hold a displayable instance and must quote it verbatim from a named file.

Pattern	Repair direction	Reported location
Cross-section references	Refer to an object from one sentence that uses it, placed where the result is interpreted or the method is applied. Leave genuinely local objects unchanged. Add no sentence that only points.	Label, object kind, home section
Macro redundancy	Rewrite the recycled sentence so that it adds a detail, consequence, or qualification its section needs, or remove it. Do not paraphrase to disguise repetition.	Section, sentence, source section
Argument graph	Move the context that leads up to a shallow claim ahead of it, or move the claim after that context. Do not add, delete, or weaken claims.	Claim sentence and the supporting sentence that currently follows it
Citation isolation	State how a cited work relates to another work already cited, in one sentence citing both, using only relations the manuscript already supports.	Section and citing sentence
Figure exposition	Remove the expository elements and state them once in the text if needed. Keep every component and connection. The editor lists the phrases to erase and a deterministic step erases them.	Figure, element type, transcribed element text
Evidence gap	Never invent an instance. **(a)** If the experiment records hold a specimen, display it verbatim with its source file. **(b)** If they hold none, say so where the aggregate result is interpreted: no individual case was retained or inspected.	One manuscript-level observation: silent, acknowledged, or exhibited

Table 10: Repair directions and location formats of `SciSlop_v0.5.md`. All six entries supply feedback and enter the stopping rule.

C.3 Review against scientific records

Overview. The reviewer reads every change of the round as a before-and-after pair and keeps or reverts each one. It checks the change against the original manuscript and the project records.

- **Changes.** The controller aligns M_{r-1} and $\widetilde{M}_r$ paragraph by paragraph. Each differing paragraph becomes a numbered change with its before-and-after text. A figure edit becomes one further change described by its edit report.
- **Review call.** Changes are reviewed three at a time, in parallel. Each call receives the rules below, its three changes, the full revised and original text of the files they touch, and the lines of the records that contain any number used in the revised text. It returns keep or revert, a category, and a reason quoting the decisive words for each change.
- **Revert rules.** The revised text says something that neither the manuscript nor the records support. A reported number is changed, rounded, dropped, or introduced without an exact match. A reference, section pointer, or citation is added to a sentence that does not use it. Two cited works are joined by a relation the manuscript never stated. A figure, table, label, citation command, or result is removed. A figure edit erased a phrase the check did not list or a component of the mechanism. A new exhibit is not an exact copy of a record with its source named. Citing sentences are merged without a supported relation.
- **Keep rules.** The change repairs a listed pattern with material already in the manuscript or records, in a sentence that uses what it refers to, and leaves every number and cited work as it was. A sentence stating that no individual case was inspected and that results rest on aggregate scores is kept when the records hold no specimen.
- **Application.** A change mixing supported and unsupported edits is reverted in full, with the reason naming the unsupported part. Only an explicit keep retains a change. Everything else is restored from M_{r-1}.
- **Retirement.** A separate call proposes instances to leave alone for the rest of the procedure. It must look for a home for the object in every other section and name the sections it checked. The controller applies retirement only to section instances and to the evidence-gap observation. Retired instances leave the feedback but stay in the measurements.

C.4 Measurement, preservation, and stopping

- **Measurement after review.** The retained manuscript is measured after review, so reverted edits cannot contribute to the reported reduction. The four deterministic measurers run every round. The argument graph is remeasured whenever the introduction changed, with the same labeling server and estimator as the benchmark. The figure exposition is rescored from the transcript of the current image. A manuscript whose measured evidence gap is 1.0 and whose text states that no individual case was retained or inspected and that the claims rest on aggregate scores is recorded as an acknowledged gap of 0.5. The matched human values are the unmodified measurer values.
- **Preservation checks.** Table 11 lists the checks computed against M_0 after every round. A hard violation disqualifies the round as successful mitigation. Audit flags mark changes that may be legitimate but require inspection, such as a removed recycled sentence that also removed a repeated number. Numbers are compared as tokens after removing comments and the arguments of citation, reference, label, input, and graphics commands.
- **Stopping.** The procedure stops when no instance remains after retirement, when the editor changes nothing, when review restores the whole round twice in a row, or after three rounds. The last retained manuscript is returned with its per-round trees, measurements, verdicts, retirements, edit reports, and preservation records.

Check	Condition to report relative to M_0	Class
Cited works	Addition or removal of a cited bibliography key	**Hard**
Dangling citations	A newly unresolved citation key	**Hard**
Reference targets	A reference to an absent label	**Hard**
Numeric values	An original body numeric token is absent	**Hard**
Body length	Revised-to-original word ratio outside $[0.8, 1.25]$	**Hard**
Labels	An original label is removed	Audit
New numbers	A numeric token is added	Audit
Numeric mentions	A token originally occurring at least twice occurs fewer times	Audit
Denominators	A deterministic pattern's score and denominator both decrease	Audit
Evidence gap	An originally silent gap is measured as acknowledged or exhibited	Audit

Table 11: Preservation checks recorded after every round. **Hard** constraints define admissible mitigation. Audit flags require examination of the underlying change.

C.5 Distance to human means

Measure	Original	Base prompting	Claude Code	Reviewer-based	Slop-aware	SciSlopHarness
Macro redundancy	+0.399	+0.314	+0.191	+0.213	−0.400	**+0.100**
Cross-section refs.	+0.201	+0.193	+0.180	+0.141	−0.660	**+0.078**
Argument graph	+0.200	+0.200	+0.200	+0.200	+0.200	**+0.048**
Citation isolation	+0.289	+0.279	+0.188	+0.290	−0.411	**+0.085**
Figure exposition	+0.265	+0.265	+0.265	+0.265	+0.265	**+0.089**
Evidence gap	+0.427	+0.432	+0.372	+0.372	+0.372	**+0.119**
Sum of absolute distances	1.780	1.682	1.395	1.480	2.307	**0.519**

Table 12: Distance from the human mean after the third revision round (revised mean − human mean) on the all papers of Figure 5. Bold marks the smallest distance per row.

Table 12 gives the values behind the 63% reduction reported in Section 5.3. For each measure, the distance is the mean score of the revised papers after the third round minus the matched human mean, the quantity plotted in Figure 5. The gap of a method is the sum of its absolute distances over the six measures. Claude Code has the smallest gap among the four baselines, so the reduction is computed against it. SciSlopHarness closes the gap from 1.395 to 0.519, a reduction of 63%, and ends closest to the human mean on every measure. Slop-aware revision passes below the human mean on Macro redundancy, Cross-section references, and Citation isolation, so its gap grows over rounds although its scores fall.

D REVISION BASELINE DETAILS

This appendix specifies the four revision baselines compared with SCISLOPHARNESS in Section 5: base prompting, Claude Code, reviewer-based refinement, and slop-aware revision. The baselines differ only in the interface through which the editor acts and in the instruction it receives. The editor model, the manuscript state, the round protocol, and the measurement are shared. Appendix C specifies SCISLOPHARNESS itself.

D.1 SHARED PROTOCOL

Editor and rounds. Every baseline uses `claude-haiku-4-5-20251001` as the editor, the same model as the default SCISLOPHARNESS editor in Table 9. Each paper is revised for three rounds. Round r starts from a copy of the manuscript retained after round $r-1$, and round 1 starts from the original AI-generated manuscript. The copied tree contains the root `.tex`, `.bib`, `.bst`, and `.sty` files and `sections/*.tex`. Figures and compiled PDFs are not copied, because the editor cannot read them and the measurers do not use them. Each round issues one editor call, which for agent conditions is one agent session that may contain several tool interactions.

Information given to the editor. No baseline receives human papers, matched human reference values, slop scores, score formulas, or thresholds. All four baselines are instructed not to fabricate experiments, numbers, or citations. The boxes below show each baseline's system and user messages, including its exact constraints. Slop-aware revision additionally receives the pattern definitions and located instances described in Appendix D.5.

Execution record. If the editor makes no changes, we retry once. If the retry also makes no changes, we exclude that round. We save each prompt and revision and evaluate revisions with the same metrics. We also check for content loss and citations missing from the bibliography.

D.2 BASE PROMPTING

Base prompting is a plain text condition without tools. The editor receives the complete LaTeX source of the paper inside the prompt and returns the complete revised source of every file inside file markers. The user instruction is the generic improvement instruction followed by the format specification.

Base prompting

System. You are an expert editor of machine-learning research papers written in LaTeX. You return revised LaTeX source only.

User. Revise this paper so that it is a better research paper: improve its clarity, organization, argumentation, and presentation, and make sure every part of the paper does its job for the reader. Keep the method, the experiments, and the reported results as they are. Do not fabricate experiments, numbers, or citations; if evidence is missing, narrow the claim instead. Do not add citations that are not already in the paper, and do not change reported numbers.

This is revision round ⟨r⟩. The complete LaTeX source of the paper follows, file by file. Return the revised source of every file, in the same order, using exactly this format and nothing else:

`<<<FILE {path}>>>`

(the complete revised content of that file)

`<<<END {path}>>>`

Return all ⟨n⟩ files (⟨paths⟩). Keep every `\input`, `\label`, `\cite`, `\ref`, environment and preamble line that a file already has, unless the revision moves or removes prose; never remove `\begin{document}`, `\end{document}`, `\documentclass`, or `\input` lines. Do not abbreviate or elide any part of a file.

⟨source of each file, preceded by `===== FILE {path} =====`⟩

A returned file replaces the original only if it is present in the output, its length lies between 0.5 and 2.5 times the original, it preserves every `\input`, `\documentclass`, `\begin{document}`, and `\end{document}` token of the original, and it is not identical to the original. A file that fails any check keeps its previous content, and the reason is logged. These checks protect the compilability of the tree and do not inspect the prose.

D.3 Claude Code

Claude Code (Anthropic, 2025) runs with its stock system prompt in the paper directory, with the same editor model as the other conditions. It runs in restricted mode, which removes shell and code execution and ignores user configuration, so the available tools are file reading, searching, and editing. The instruction is the generic improvement instruction of base prompting with one added sentence.

Claude Code

System. *The stock Claude Code system prompt is used without a custom replacement.*

User. Revise this paper so that it is a better research paper: improve its clarity, organization, argumentation, and presentation, and make sure every part of the paper does its job for the reader. Keep the method, the experiments, and the reported results as they are. Do not fabricate experiments, numbers, or citations; if evidence is missing, narrow the claim instead. Do not add citations that are not already in the paper, and do not change reported numbers.

Make the changes directly by editing the .tex files in the current directory; do not just give advice.

D.4 Reviewer-based refinement

Reviewer-based refinement alternates an automated review with an editing session. At each round, Qwen2.5-32B-Instruct (Yang et al., 2024) reads the current manuscript source and returns a structured review. The review is regenerated from the current manuscript each round, so later rounds respond to the revised paper rather than to the original. The source is passed as text truncated to 50,000 characters. If the model returns no valid JSON, the call is repeated at 24,000 and then 14,000 characters.

Reviewer-based refinement / Review generation

System. *No separate system message is supplied.*

User. You are a knowledgeable, critical, but fair reviewer for a top machine-learning conference. Read the paper draft below and write a review of its overall quality: the significance of the problem, the soundness of the approach, the strength of the evidence, and the clarity of the writing.

Respond with ONLY compact JSON in exactly this format:

```
{"summary": "<2-4 sentence summary of the paper and its contributions>",
 "strengths": ["<strength 1>", "<strength 2>", "<strength 3>"],
 "weaknesses": ["<weakness 1>", "<weakness 2>", "<weakness 3>"],
 "questions": ["<question 1>", "<question 2>", "<question 3>"],
 "rating": <integer 1-10, where 1 = trivial or wrong and 10 = seminal>}
```

PAPER (LaTeX source, possibly truncated):

⟨manuscript source⟩

The editor is an agent with a neutral system prompt and with read, edit, and write tools. Web search is denied and shell commands are not permitted. The review prompt and the editing instruction are general. Neither mentions SciSlop, the six patterns, or any measured quantity.

Reviewer-based refinement / Manuscript editing

System. You are revising a LaTeX research paper. Source files (main.tex, sections/*.tex, math_commands.tex, a .bib file) are in your working directory. Use Read/Edit/Write to modify the .tex files in place. Do not create files outside the paper directory. When done, finish your turn; do not run latex and do not summarize.

User. A reviewer left the following review of this paper. Revise the paper to address the review.

⟨review JSON⟩

Do not fabricate experiments, numbers, or citations; if evidence is missing, narrow the claim instead.

D.5 SLOP-AWARE REVISION

Slop-aware revision gives the editor a list of detected slop problems and instructions for fixing them in place of a general review. The editor interface, system prompt, tools, and budget are the same as in reviewer-based refinement. At the start of each round, we check the current manuscript and prepare one prompt containing the detected problems. For each problem, the prompt gives its name, definition, repair instructions, and locations in the manuscript. The definitions and instructions follow the measurer specifications. We check the first four patterns of Table 13, adding *Argument graph* and *Figure exposition* for the comparison with SCISLOPHARNESS. The prompt lists up to 12 instances per pattern and reports the number of additional instances. For *Evidence gap*, it simply states that the manuscript contains no concrete example. Only detected problems are included, and a round is skipped if no problems are found. The editor receives no scoring formulas or target scores. Table 13 gives the definition and repair text of each pattern.

Slop-aware revision

System. You are revising a LaTeX research paper. Source files (main.tex, sections/*.tex, math_commands.tex, a .bib file) are in your working directory. Use Read/Edit/Write to modify the .tex files in place. Do not create files outside the paper directory. When done, finish your turn; do not run latex and do not summarize.

User. An automated check of this manuscript found the following issues. Revise the paper to address them. Do not fabricate experiments, numbers, or citations; if evidence is missing, narrow the claim instead. Do not add citations that are not already in the paper, and do not change reported numbers. Edit the .tex files in place.

This is revision round $\langle r \rangle$.

Issue 1: ⟨pattern name⟩

⟨definition⟩

What to do: ⟨repair direction⟩

Findings ($\langle k \rangle$ total; ⟨unit label⟩):

- ⟨located unit⟩ (up to 12 lines)

- (and $\langle k - 12 \rangle$ more of the same kind; treat them the same way)

Issue 2: ...

Pattern	Definition	What to do
Cross-section references	A paper declares objects: its body sections and the labelled figures, tables, equations, and algorithms inside them. An object is unused when no section other than the one that contains it ever refers to it, so the rest of the paper never builds on it.	Where the paper's argument actually relies on an object, refer to it from the section that relies on it (for example, discuss the result of a table in the section whose claim it supports, and point back to the method's equation where the experiments use it). Do not add pointers that the surrounding text does not use.
Macro redundancy	A sentence is recycled when at least half of it repeats, almost word for word, text that an earlier, different section already contained.	Rewrite each recycled sentence so that it adds what its own section needs (a new detail, a consequence, a qualification), or remove it if the section does not need it. Do not paraphrase merely to disguise the repetition.
Citation isolation	A citing sentence is isolated when it cites a single prior work and neither groups it with another work, nor mentions another work elsewhere in the sentence, nor states a relation between two works. A related-work section made of isolated sentences lists prior work instead of positioning it.	Where the paper's positioning depends on a cited work, state how it relates to other works already cited in the paper: a shared limitation, a difference in assumption or method, or what this paper takes from each. Only use works already cited in the paper; do not add new citations.
Evidence gap	The paper reports aggregate results but never displays a single concrete instance of what it counts: an example input and output, a case, a failure, a worked example, or a quoted specimen.	If the paper's own materials in this directory contain such an instance (for example a prompt, a generated output, a case in the appendix), display one where the reader needs it and refer to it from the text. If no instance is available in the materials, do not invent one; instead say explicitly, in the relevant section, that no instance is shown.
Argument graph	A key claim of the Introduction is a sentence asserting superiority, a prior limitation, or a design choice. The claim's strongest contextual cue is the Introduction sentence that most raises the claim's plausibility. The claim is declared rather than argued when that cue occurs after the claim, so the Introduction states the conclusion before building toward it.	Reorder so that the context precedes the claim it supports. Move the limitation of prior work, the observation, or the design constraint ahead of the sentence that asserts the contribution, or move the assertion down to follow it. Do not delete claims and do not add new claims. Do not weaken a claim to avoid arguing for it.
Figure exposition	A method diagram should depict the mechanism. Its components and how they connect. A box, panel, or caption element is expository when it carries material that belongs to the text rather than the mechanism. Experimental settings, dataset names, numeric results, interpretive claims, a legend that restates the prose, or a step-by-step narration of the pipeline.	Remove the expository material from the method diagram and, where the reader needs it, state it once in the text near the figure. Keep only components of the method and their connections. Remove exactly the elements listed as expository and keep every other element and every arrow the image has. Do not invent components the figure does not show, do not simplify the mechanism to make the drawing easier, and do not drop the figure.

Table 13: Definitions and repair instructions for slop-aware revision. The first four rows are used in the four-pattern setting.

E ICLR DATA AND MEASUREMENT

Figure 4 uses two corpora of human-written ICLR submissions, 1,177 papers from 2017 to 2025 and 691 from ICLR 2026, scored by every system of Table 2 with the same scripts, settings, and document views. The measurers and the calibration were fixed on SCISLOPBENCH, and every statistic is computed within a year or within ICLR 2026, because the ICLR rating scale changed over the years.

Corpora. ICLR is the one venue where every submission, accepted or rejected, carries public review scores and a decision, and its arXiv LaTeX e-prints give the source that the measures read. The multi-year corpus therefore holds ICLR 2017 to 2025 submissions with both, nine years that span the arrival of LLM writing. It was collected in strata of year by integer rating and of year by award, so that every year holds papers from rating 2 to rating 9 and its orals; at the ends of the scale the strata take every paper that exists. Of 1,381 decided papers, 204 fail the extraction rules below, leaving 1,177, 590 rejected, 434 accepted, and 153 oral, with 78 to 240 papers a year. The ICLR 2026 corpus serves the review dimensions. It is the one year with a soundness, presentation, and contribution score for every paper on one rating scale, and the one year with an independent commercial AI judgment of the same full text, an analysis that Pangram published for every ICLR 2026 submission. Its 691 submissions with an arXiv e-print were sampled over six bands of that AI share crossed with the decision, so that the high-AI tail, mostly rejected and rarely on arXiv, is present; 560 are accepted and 131 rejected.

Exclusion. A paper is removed only when its source cannot be measured: a missing `\input` file, a main file that is a supplement, a failed section parse, fewer than three body sections, fewer than five declared objects, a prose view under 1,500 words, or any measure below its minimum denominator (Appendix A). ICLR 2026 adds fewer than three reviews and a body length outside the 1st to 99th percentile. No rule reads a score or a rating, and a removed paper is removed for every system.

Systems and scores. Detectors read the prose view, reviewers the body source, and the measures the full source. Every score is oriented so that a higher value means more AI-like, which reverses Binoculars and the reviewers' Overall rating. The detectors score all 1,177 papers; the two reviewer models score a subset drawn by rating quintile within each year and filled at random by year and decision, 287 papers for CycleReviewer and 282 for AI Scientist. The SciSlop line is the aggregate of Eq. 7 (Appendix A) scored on every paper of both corpora. *Argument graph* was not run on ICLR 2026, and *Evidence gap* and *Figure exposition* apply to 869 and 662 of the multi-year papers.

Panel (a). Each system is calibrated once on the FARS pairs of SCISLOPBENCH by a logistic regression of the AI label on its raw score, and the fitted curve is applied unchanged to the ICLR papers, giving the probability the system assigns to AI authorship. Each probability is adjusted for year by subtracting the mean of its own year and adding back the mean over all years, so the axis stays on the probability scale; a paper sits at its rounded mean rating. Levels 2 to 9 hold 31 to 217 papers each; levels 1 and 10, with 11 and 1 papers, are not drawn.

Panel (b). A cell is the Spearman correlation between a system's oriented score and the mean reviewer score on one dimension of ICLR 2026, over the papers the system scored, 686 for the detectors, 691 for Pangram, 680 for SciSlop, and 338 and 285 for CycleReviewer and AI Scientist.

Panel (c). Within each year, the value is the probability that a rejected paper scores more AI-like than an accepted or oral paper of the same year. A system is drawn when at least 15 papers stand on each side in at least five years; the reviewer models meet this in three years and are not drawn.

F ADDITIONAL MATCHED REVISION CASES

Section 5.4 compared slop-aware revision and SCISLOPHARNESS on the same passages for two measures, Macro redundancy and Cross-section references, and showed that the two lower the score by different means, one by deleting or scattering information and the other by repairing the located passage. This appendix extends that comparison to the remaining four measures, Citation isolation, Evidence gap, Argument graph, and Figure exposition.

The same distinction holds for the remaining four measures. Table 14 repeats the format of Table 4 for the four measures it does not cover and shows the same pattern. Slop-aware revision lowers the score by leaving, deleting, or inventing content, whereas SCISLOPHARNESS repairs the located passage itself.

- **Citation isolation** shows whether a repair relates the cited work to anything else. Slop-aware revision rewrites the CALM sentence but still describes CALM alone, so the located sentence stays isolated; its lower manuscript score comes from changes elsewhere in the paper. SCISLOPHARNESS keeps the sentence and adds the contrast with InfMem, which stops at the chunk level and reports a 3.3–5.1× speedup. The isolation is removed by adding a relation, not by rewording the sentence.
- **Evidence gap** shows what a revision does when the record holds no concrete instance. Slop-aware revision reaches a zero score by inserting a worked example whose question, responses, and numbers appear nowhere in the manuscript or the project records. SCISLOPHARNESS states that the claims rest on aggregate results only, and its review reverts the one example the editor inserted as beyond the record. The gap stays acknowledged at 0.5 instead of being filled with an invented instance.
- **Argument graph** shows whether a claim follows the method it depends on. Slop-aware revision leaves the design claim ahead of the method and adds a section pointer instead. SCISLOPHARNESS moves the claim directly after the method description without deleting either sentence, and the manuscript score falls to zero. The repair reorders existing content rather than adding text.
- **Figure exposition** shows whether a revision can reach text inside an image. Slop-aware revision edits source text only, so all eight phrases that restate the paper remain in the diagram and the score stays at its original value. SCISLOPHARNESS erases six of the eight phrases while leaving the components and arrows intact. This comparison is not under equal tools, since the baseline has no image editor, but it shows how far the harness's image edits reach.

Citation isolation		Shaded sentence relates CALM to nothing. × Slop-aware leaves CALM alone. ✓ SCISLOPHARNESS adds InfMem.
Original		For language models, CALM [1] uses confidence-based early exiting at the token level, achieving significant speedups on text generation.
Slop-aware	×	For language models, confidence-based methods like CALM [1] extend these layer-wise approaches to the token level, achieving significant speedups during text generation…
SCISLOPHARNESS	✓	For language models, CALM [1] uses confidence-based early exiting at the token level, achieving significant speedups on text generation, whereas InfMem [2] learns stopping policies at the chunk level for memory agents, achieving 3.3–5.1× speedup.
Evidence gap		No concrete instance appears in the paper. × Slop-aware invents one. ✓ SCISLOPHARNESS states the gap.
Original		*No example input, output, case, or failure is displayed; the paper reports aggregate accuracies and agreement rates only.*
Slop-aware	×	Worked Example. Consider a physical reasoning question: "A rubber ball is dropped from rest and falls 5 meters…" Response A (preferred): "The ball accelerates downward at approximately 9.8 m/s^2 due to gravity…"… *Nothing in it occurs in the manuscript or the project records.*
SCISLOPHARNESS	✓	This manuscript reports aggregate results from experiments but does not display concrete individual instances such as example inputs, outputs, or specific failure cases. The claims regarding model behavior and consistency gating effects rest on aggregate scores only. *Review reverts the editor's own inserted example* (e.g., consistent selection of option A) *as beyond the record.*
Argument graph		Shaded claim precedes its method. × Slop-aware keeps the order. ✓ SCISLOPHARNESS moves it after the method.
Original		This runtime feedback can guide diffusion models to target repair at semantically relevant regions… ↓ We introduce an execution-trace guided remasking pipeline that uses runtime diagnostics… to localize failures and target repair in diffusion code generation.
Slop-aware	×	This runtime feedback can guide diffusion models to target repair at semantically relevant regions… ↓ We introduce an execution-trace guided remasking pipeline described in Section [Method]… to localize failures and target repair in diffusion code generation.
SCISLOPHARNESS	✓	We introduce an execution-trace guided remasking pipeline that uses runtime diagnostics… to localize failures and target repair in diffusion code generation. This runtime feedback can guide diffusion models to target repair at semantically relevant regions…
Figure exposition		Shaded diagram text restates the paper. × Slop-aware leaves the image. ✓ SCISLOPHARNESS erases six of eight phrases.
Original		Stage 1: Rollout Generation · Stage 2: Metamorphic Audit · Stage 3: Trust Scoring · [agent icon] = Agent; × = Outcome · Legend: · Orange box = State-preserving command · Purple box = Distance computation & Scoring · Green box = World Model
Slop-aware	×	*The baseline edits source text only, so every phrase above remains in the image.*
SCISLOPHARNESS	✓	Stage 1: Rollout Generation · Stage 2: Metamorphic Audit · Stage 3: Trust Scoring · [agent icon] = Agent; × = Outcome · Legend: · Orange box = State-preserving command · Purple box = Distance computation & Scoring · Green box = World Model *Green phrases are erased in round 1 and review confirms that nothing outside their rectangles changed; red phrases are not located and remain.*

Table 14: One passage per remaining pattern before revision, after slop-aware revision, and after SCISLOPHARNESS, marking the located slop, what a revision lost or left, and what it kept or repaired. Brackets replace citation and reference commands, and the arrow marks the order of two statements across intervening text. Manuscript scores (original / slop-aware / SCISLOPHARNESS): 0.615 / 0.385 / 0.571, 1.000 / 0.000 / 0.500, 0.333 / 0.167 / 0.000, and 0.500 / 0.500 / 0.167, where the baseline edits no images and so keeps the original figure score.