

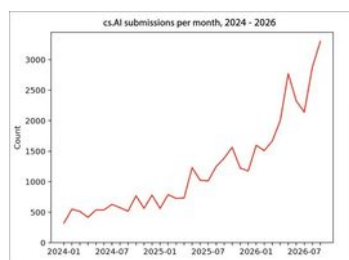
Fighting AI Slop in Science Papers

Originally published Sunday 4th October 2026 at <https://www.unite.ai/fighting-ai-slop-in-science-papers/>



AI does everything at scale, from [DOS-attack-level web scraping](#) to producing, or enabling, [such vast volumes](#) of scientific research submissions that the result is becoming both a logistical crisis and a [crisis of quality](#), as the signal-to-noise ratio continues to become unnavigable.

Last week, preprint server arXiv [announced](#) that it will be instituting a new rate-limiting policy for submitters, who will now be capped at two submissions per month. The measure has been taken, the announcement suggests, in the face of a vertiginous rise in submission rates over a short time-period:



Monthly submissions to arXiv's cs.AI category from January 2024 to September 2026, showing a more than sixfold increase over the period. [Source](#)

Kat Boboris's blog post announcing the move, which [drew comment](#) at Hacker News last Thursday, stated that arXiv received 40,363 submissions in September 2026 – almost double the 20,569 received in September 2024, and more than four times the 9,869 received in September 2016.

The latest month's submissions, Boboris observed, also generated almost 9,000 support tickets for arXiv staff and moderators:

'Our moderators are observing an increase in thin papers of narrow scope, as well as 'salami' papers, where a single work is broken up and submitted as a set of smaller papers.'

'There is also a marked increase in dense, AI-written papers. AI tools are making it easy for authors to flood arXiv and other repositories with these low-value papers.'

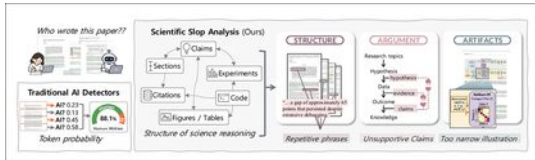
Already, in May of this year, arXiv had taken the measure of implementing a [one-year ban](#) for unchecked AI content; and in October of last year, had told submitters of survey papers and position papers (both of which can be generated with less effort than a full academic study) [that they would need peer-backing](#) in order to be published at arXiv.

For the moment, the arXiv domain's often obstructive capping of http requests is not too much in evidence, and one can only hope that its very useful RSS feeds survive this ongoing retrenchment. Last week Reddit [announced](#) the long-feared total elimination of its RSS feeds, which will take place from the middle of next month. However, arXiv's non-profit status means that there's little [similar capital](#) to be gained by shepherding readers into mandatory site visits, or [enforced logins](#) – at least, for the moment.

Against the Rising Tide

In the face of such [severe and growing problems](#) around the negative effect of AI use in science research – most especially regarding AI-related research, which has [eclipsed](#) all other categories, and risen from obscurity to become [a political and economic signifier](#), lately – a strand of research has emerged examining ways to counter the decline in quality of AI-related paper submissions.

The latest to address the problem comes in the form of a collaboration between Korea's Seoul National University and the University of Minnesota in the US. The [paper](#), titled *Science or Slop?: Benchmarking and Mitigating Scientific Slop in AI-Generated Papers*, proposes measuring 'scientific slop' through failures in the connections between a paper's claims, evidence, citations and structure:



From the new paper, an overview of the work's approach to 'scientific slop', showing how failures in structure, argument and supporting artifacts can reveal weaknesses that conventional AI-text detectors miss. [Source](#)

Unlike prior approaches, the new system looks beyond the text itself to assess whether the paper's claims are properly supported by its arguments, evidence and citations. The authors state*:

'Each part of a paper can look plausible in isolation, so such breakdowns 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](#), repairing these patterns requires restoring the missing relations without changing the underlying science.'

The authors' new benchmark, dubbed *SciSlopBench*, has been embodied into *SciSlopHarness*, a framework designed to detect and repair failures in a paper's scientific reasoning, while preserving the underlying scientific evidence:

Evidence gap	No concrete instance appears in the paper. × Slop-aware invents one. ✓ SCISLOPHARNES 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.
SCISLOPHARNES	✓ 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. ✓ SCISLOPHARNES 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.
SCISLOPHARNES	✓ 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...

Examples comparing flawed passages with revisions made by *SciSlopHarness*. Unsupported additions are rejected, while claims are reordered or rewritten where evidence available in the paper.

The *SciSlopBench* benchmark itself was built from 390 AI-generated papers, each paired with a human-written paper addressing a similar research problem and contribution type.

In tests, *SciSlopBench* was used to distinguish between 390 pairs of papers, with each pair consisting of the aforementioned AI-generated paper, and a human-written paper matched for research problem and type of contribution. The benchmark correctly identified the AI-generated paper in 85.9% of these comparisons, substantially outperforming conventional AI-text detectors.

The authors state:

'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.'

In addition to the contributions made by the paper itself, the authors have operationalized the principles of their new work in the form of a [live demo](#) where users can submit scientific papers for analysis of the amount of AI slop they contain.

Interestingly, the new paper itself, analyzed by the demo, comes in at [a 'moderate' slop score](#) of 40/100[†]:

The new paper, analyzed by its own algorithm, gets 'slop' areas flagged by the authors' new process. [Source](#)

Method and Data Approach

The 2025 collaboration [Why Slop Matters](#) described 'superficial competence' as a defining characteristic of AI slop, where apparent quality conceals a lack of substance. The new work applies this idea more specifically to scientific papers, defining 'scientific slop' as failures that make it difficult to follow how a study is organized; how its claims are supported; and how its methods and evidence can be examined, with the failures grouped into *structure*; *argument*; and *artifacts*:

Item	Illustration	Unit of analysis $U_j(p)$	Score $s_j(p)$	What $s_j = 1$ means
Cross-section references (STRUCTURE)		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 graph (ARGUMENT)		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.
Figure exposition (ARTIFACTS)		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.

Measurement rules for the six types of scientific slop used in the benchmark. The table shows what is examined for each measure, how the score is calculated, and represents, with higher scores indicating more extensive failures.

The six measures of scientific slop defined in the paper are *cross-section references* (whether sections refer meaningfully to material elsewhere in the paper); *macro redundancy* (whether later sections repeat earlier material); *argument graph* (whether claims are properly supported by preceding reasoning); *citation isolation* (whether citations are used superficially, without explaining how the cited works relate to the paper or to each other); *figure exposition* (whether method figures actually explain the method); and *evidence gap* (whether reported results are supported by concrete examples).

The benchmark was constructed by pairing AI-generated papers with comparable/equivalent human-written papers, with the AI papers curated from [FARS](#) and the [2025 Agents4Science](#) competition.

For each machine-generated FARS paper, the researchers searched its citations for a human-authored paper of the same type that had been accepted at a top-tier venue, and then selected the closest match by research topic, producing 143 pairs. The Agents4Science papers were likewise paired with human-written counterparts, producing a further 247 pairs, and bringing the benchmark to 390 AI-human pairs in total. The papers span the life, social and natural sciences, though the dataset is weighted heavily towards computer science.

For metrics, performance was evaluated using [PairAcc](#), which measures how often the human paper is ranked above its AI counterpart; and [AUROC](#), which measures overall separation between human and AI papers across different thresholds. The authors also report detection performance when the false-positive rate for human papers is fixed at 5%.

Tests

Initial tests compared SciSlop against the AI-text detectors [Binoculars](#); [DetectGPT](#); and [NTS^{††}](#), as well as the automated review systems [AI Scientist](#) Reviewer and [CycleReviewer](#).

For the later revision tests, four baselines were used: *base prompting*; *Claude Code*; *reviewer-based refinement*; and *slop-aware revision*. The first three received only general instructions to improve the paper, while slop-aware revision was additionally given the definitions and locations of detected slop. All were compared with SciSlopHarness under equivalent revision conditions.

For implementation, the text detectors were given the paper’s prose, while the other methods received the paper source and, where required, access to its code. Binoculars used [Falcon-7B](#) base and instruct models; DetectGPT used [T5-3B](#) to generate 100 perturbations; AI Scientist used [Qwen2.5-32B-Instruct](#); and CycleReviewer used its released 8B checkpoint. For papers exceeding Falcon’s 2,048-token limit, non-overlapping windows were scored separately and averaged.

The initial detection results are detailed in the two tables below: SciSlop achieved 85.9% PairAcc, compared with 68.7% for Binoculars and 68.5% for the strongest automated reviewer, cutting the error rate by more than half:

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

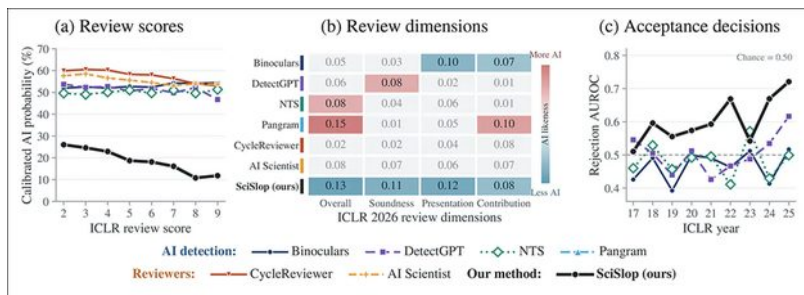
Plane	Measure	PairAcc	AUROC	TPR
(STRUCTURE)	Macro redundancy	0.723	0.727	0.467
	Cross-section refs.	0.905	0.895	0.648
(ARGUMENT)	Argument graph†	0.586	0.584	0.159
	Citation isolation	0.793	0.782	0.444
(ARTIFACTS)	Figure exposition†	0.809	0.786	0.364
	Evidence gap	0.764	0.765	0.000
SciSlop (ours), mean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

Detection results comparing SciSlop with AI-text detectors and automated reviewers, alongside results for its six individual slop measures. SciSlop achieves the highest cross-section references provide the strongest individual result.

Cross-section references performed particularly strongly on their own, reaching 90.5% PairAcc without requiring a model, and detecting 65% of AI-generated papers at a 5% false-positive rate, compared with 24% for Binoculars.

The authors argue that these results indicate that relationships across the paper provide a stronger signal of AI generation than conventional text-level detection on this dataset, while also identifying specific weaknesses that could subsequently be targeted for revision.

The relationship between scientific slop and human assessments of paper quality was examined next. As shown in the first column below, higher slop was associated with lower ICLR scores for overall rating, soundness, presentation and contribution:

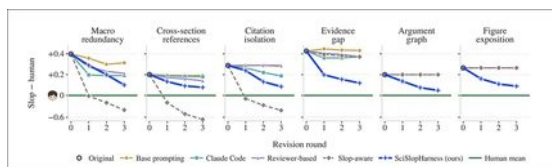


Comparison of SciSlop with AI-text detectors and automated reviewers against ICLR review outcomes. The left panel shows AI-likeness against review scores; the review dimensions; the right shows performance in distinguishing rejected from accepted papers across nine years.

Rejected and accepted papers were also distinguished above chance across all nine years examined, while conventional detectors fell below chance in most comparisons.

Scientific slop was therefore found to reflect weaknesses already penalized by human reviewers, rather than functioning solely as a signal of AI authorship.

The final tests examined whether SciSlopHarness could remove slop that remained after more general revision. As shown below, the harness finished closest to the human average across all six measures, while general revision often left substantial slop and direct slop-aware revision sometimes overcorrected:



Revision results comparing SciSlopHarness with four baseline methods across the six slop measures. Values closer to zero are closer to the human-paper average, with SciSlopHarness generally moving towards this level while slop-aware revision frequently overshoots it.

Each proposed change was checked against the paper’s scientific reasoning and supporting evidence, with unsupported edits rejected. Across the six measures, the remaining gap from human-written papers was reduced by 63% compared with Claude Code, the strongest revision baseline.

Conclusion

It was interesting, in the course of writing this piece, to note not only how poorly this paper scored on its own demo site, but to observe at least one example of a ‘lazy’ or ‘cosmetic’ citation^{††}, which has lately become a minor but growing curse in research papers.

In such cases, beyond this particular paper, researchers seem to draw on their own knowledge rather than engaging meaningfully with the existing literature. Yet, constrained by convention to 'show their work', citations are often supplied with what reads as impatience, even contempt for the process, and often apparently relies on the reader to accept a glut of references as a sufficient 'patina' of provenance, albeit one that would not survive excessive scrutiny.

Arxiv is fighting back against the AI-driven industrialization of submissions; whether there will be similar constraints on AI's direct involvement in research, absent some related disaster of sufficient magnitude, remains to be seen.

** The authors' emphases, not mine – but my conversion where necessary of the authors' inline citations to hyperlinks.*

† The authors do not claim zero AI use in their own paper, and [detail such usage](#).

†† It may interest the reader to know that I had to hunt down a correct link myself for the NTS framework, because the link supplied by the authors was not relevant – one of the very metrics that SciSlop keys on!

First published Sunday, October 4th, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Website: martinanderson.ai

Contact: martin@martinanderson.ai