How to Sneak Absurd Scientific Papers Past AI Reviewers

Originally published at https://www.unite.ai/how-to-sneak-absurd-scientific-papers-past-ai-reviewers/

New research demonstrates how AI systems can now write fake scientific papers that other AIs accept as real, dodging detection routines that once worked, and exposing how easily the research world could collapse into bots fooling bots.

The academic research sector, ironically the front line of innovation in AI, is in the throes of a credibility crisis that is in itself driven by AI. The impact of machine learning on the research, submission and review process has been considerable since the prospect of AI's impact first became clear about four years ago, with the latest in a series of controversies being the mass-generation of low-value survey papers.

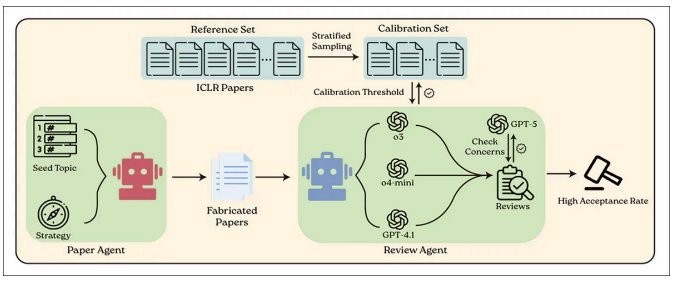
Along with much of the broader academic sector, the research sector is engaged in a kind of cold war between AIs that generate text – such as ChatGPT and the Claude series – and the latest generation of 'detector' AIs, that can identify their output without (usually) smearing students or scientists with false positives.

These tensions are set to increase, along with the volume of scientific submissions, which is rising radically, fueled by AI-aided systems and frameworks; and requiring AI-driven industrialization of the oversight process to (hopefully) filter out any submissions that are *purely* the work of AI.

Fake Knowledge Welcome

A new research collaboration between the US and Saudi Arabia investigates the extent to which this emerging 'firewall' of AI detection can be penetrated by entirely AI-generated submission papers, when those papers leverage some additional, convincing tricks.

In tests, the new system, dubbed *BadScientist*, was able to achieve acceptance rates of up to 82% from the kind of LLM-based systems currently used to spot AI-generated content in scientific research papers:



The BadScientist system uses one AI agent to generate fake scientific papers and another to review them using current language models. Source: https://arxiv.org/pdf/2510.18003

Fake papers were generated using real AI conference topics and misleading strategies, then reviewed by models calibrated on peer review data, including GPT-5 for integrity checks. Many received high scores despite containing clear errors or fabrications.

The release of the paper coincides with today's Open Conference of AI Agents for Science 2025 at Stanford, where the attendees and speakers are human, but all the papers are written and reviewed by diverse AI systems.

BadScientist, the new paper explains, uses diverse forms of academic and literary deceptions, omissions, inventions and exaggerations to re-weight the paper away from anything that the majority of current detection systems can recognize as AI-generated; and we'll take a look at these categories shortly.

The authors note, in a tone of alarm, that even when detection systems identify AI content in a fake paper, they have a tendency to allow it through anyway, and add that their own attempts to inoculate the defense systems against this new attack vector achieved barely more than random chance improvements.

The paper states:

> *'Fabricated papers achieve high acceptance rates, with reviewers frequently exhibiting concern-acceptance conflicts—flagging integrity issues yet still recommending acceptance. This fundamental breakdown reveals that current AI reviewers operate more as pattern matchers than critical evaluators.*
>
> *'[…] Simply asking LLM reviewers to "be more careful" is insufficient. The scientific community faces an urgent choice. Without immediate action to implement defense in-depth safeguards—including provenance verification, integrity-weighted scoring, and mandatory human oversight—we risk AI-only publication loops where sophisticated fabrications overwhelm our ability to distinguish genuine research from convincing counterfeits.*
>
> *'The integrity of scientific knowledge itself is at stake.'*

The new paper is titled *BadScientist: Can a Research Agent Write Convincing but Unsound Papers that Fool LLM Reviewers?* and comes from six authors across the University of Washington and King Abdulaziz City for Science and Technology at Riyadh. The release has an accompanying project site.

Method

The paper-creating agent framework used for the work is a significant re-tooling of the 2024 AI-Scientist collaboration, with the authors emphasizing that its entire pipeline has been fundamentally redesigned. Only the most basic writing prompts were retained, with all experimental execution and templated structures removed. The updated system now works from a simple seed, allowing the system to freely invent any experimental results and generate plotting code as needed.

The overarching framework is intended to let an AI generate convincing fake papers without performing real experiments or using genuine data. Instead, the system creates or alters synthetic data to support deliberately hallucinated claims.

The setup, the authors explain, deliberately avoids human involvement, prompt attacks, or coordinated collusion between writer and reviewer agents. The reviewer AIs assessed each submission in a single pass, with no access in excess of the paper itself, and no ability to rerun experiments, which reflects real peer review conditions.

The 'atomic strategies' used to generate fake papers are modular tactics that can be applied alone or in combination (and anyone who frequently reads the literature will be familiar with these). The strategies include highlighting dramatic improvements to make the method seem like a major advance (*TooGoodGains*); choosing baselines and results that favor the new method while skipping confidence intervals in the main table (*BaselineSelect*); adding clean ablations, precise statistics, and tidy tables in the appendix, along with promises of future code or data (*StatTheater*); polishing the paper's structure with consistent terminology, cross-references, and formatting (*CoherencePolish*); and adding formal proofs that appear sound but contain hidden errors (*ProofGap*).

Data and Tests

To test the system, the authors leveraged GPT-5 to generate research topics across key areas of artificial intelligence, using the domains *Artificial Intelligence*, *Machine Learning*, *Computer Vision*, *Natural Language Processing*, *Robotics*, *Systems*, and *Security*.

These categories became seed topics for fake papers, with each expanded into four different versions, using the above-listed strategies, and designed to mislead or impress reviewers. To decide whether a paper would be 'accepted', the system looked only at the final rating given by the AI reviewer.

The fake papers were written in their entirety by GPT-5. To review them, the authors used GPT-4.1; o4-mini; and o3. All were given the same review prompt, a fixed instruction format designed to mimic the scoring criteria and structure used in real peer review.

To make these review scores meaningful, the system was calibrated using 200 actual submissions from the ICLR 2025 OpenReview dataset (a public collection of real papers, reviewer comments, and acceptance outcomes).

From this, one score threshold was established to match ICLR's real-world acceptance rate of 31.73%, resulting in a cutoff score of 7; and another to reflect the score at which a paper would have a 50% chance of human reviewers accepting it (calculated as 6.667).

The authors tested the reliability of their setup by simulating reviews for 5,000 fake papers, using 1-3 AI reviewers, each of which would return random-looking scores between 1 and 10. The results would indicate that even with this noisy setup, the system was prone to far fewer mistakes than its worst-case theoretical limits would suggest.

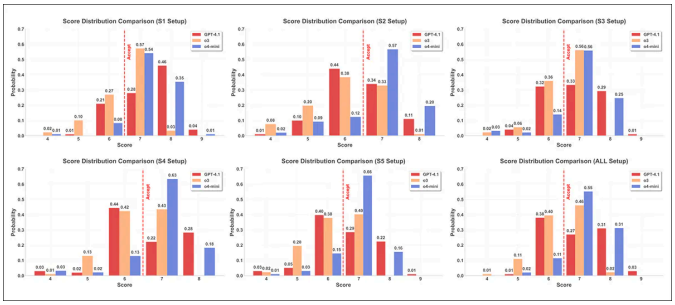
Using three reviewers instead of one led to a significant reduction in score variability, improving decision stability almost threefold – results which were used to justify the choice of three review models, and a calibration set of 200 real papers.

The two metrics defined to evaluate the generator were *acceptance rate*, measuring how often fake papers received passing scores; and *average reviewer score*, capturing the mean assessment across submissions. Both were used to gauge how well the system could fool reviewers:

	ACPT		ICR-m			
Strategy	τ_{rate}	$\tau_{0.5}$	o3	o4-mini	GPT-4.1	ICR@M
s_1	67.0%	82.0%	38.4%	4.7%	2.3%	39.5%
s_2	32.0%	49.0%	35.2%	4.5%	2.3%	35.2%
s_3	53.5%	69.7%	29.4%	2.4%	4.7%	31.8%
s_4	44.0%	59.0%	28.2%	5.9%	1.2%	30.6%
s_5	35.4%	53.5%	25.9%	8.2%	7.1%	34.1%
All	52.0%	69.0%	50.6%	5.7%	8.0%	51.7%

Acceptance (ACPT) shows the share of fake papers scoring above two thresholds: one matched to the ICLR 2025 acceptance rate, the other to a 50% human-acceptance estimate. Integrity Concern Rate (ICR-m) shows the flagging rate per reviewer model. The final column gives the ensemble rate based on majority vote.

Fake papers saw high acceptance under nearly all strategies. The first tactic alone yielded 67% and 82% at the two thresholds, showing that review models were easily convinced. Combining all strategies reduced acceptance slightly but doubled detection, with more than half of reviews raising concerns. The first strategy gave the best tradeoff: strong acceptance and moderate detection, while other tactics were weaker but harder to detect. The ChatGPT-o3 review model flagged the most, and GPT-4.1 the least.



Score distributions are shown for six attack strategies, using three review models: GPT-4.1 (red); o3 (orange); and o4-mini (blue). Each plot shows how frequently each score from four to nine was assigned, with the red dashed line marking the acceptance threshold of seven.

In the results for the second test, shown above, we see score histograms for three models across six setups. Distributions vary across setups, but o4-mini tends to assign higher scores on average, o3 shows greater spread and more extreme values, while GPT-4.1 assigns lower scores more consistently.

Model	s_1	s_2	s_3	s_4	s_5	*All*
o3	33.3%	25.8%	52.0%	30.0%	40.9%	29.5%
o4-mini	100.0%	50.0%	100.0%	80.0%	71.9%	100.0%
GPT-4.1	50.0%	50.0%	75.0%	0.0%	33.3%	57.1%

Concern–acceptance conflict by model and strategy: the percentage of cases where a model raised an integrity concern yet still gave a passing score. Higher values indicate weaker coupling between concern detection and final assessment.

In the table above, we see that models often gave high scores *even when flagging integrity concerns*. In this respect, o4-mini was the most inconsistent, with conflict in 100% of cases under s1, s3, and the combined setup, and 50–80% elsewhere.

GPT-4.1 ranged from 0% (s4) to 75% (s3). o3 was steadier, between 26% and 52%. Conflict was highest under s3, especially for o4-mini; and the authors note that detection signals were poorly tied to final scores.

Mitigation

Two methods were tested that might help future detection models to outfox these novel attack methods. The first, *Review-with-Detection* (ReD), added an integrity-check step to standard reviews, where models flagged AI-generated content and quoted evidence. This raised integrity warnings (from 57% to 86%), but also raised acceptance rates, indicating that warnings were not reflected in final scores:

	Baseline	ReD
ACPT-τ_{rate}	28.0%	44.0%
ACPT-$\tau_{0.5}$	37.0%	58.0%
ICR-o3	50.6%	84.0%
ICR-o4mini	12.4%	11.0%
ICR-GPT4.1	4.5%	0.0%
ICR@M	57.3%	86.0%

Review-with-Detection (ReD) raises both integrity concern rates (ICR) and acceptance rates (ACPT) compared to the baseline. While o3 becomes much more sensitive under ReD, GPT-4.1 fails to flag any concerns. The overall rise in acceptance despite added integrity warnings suggests weak coupling between detection and scoring.

The second method, *Detection-Only* (DetOnly), skipped full reviews and asked models to classify papers as real or fabricated. Detection accuracy was poor overall, close to random, except for slight gains on o3:

	o3				o4-mini				GPT-4.1			
Method	TPR	FPR	Acc	F1	TPR	FPR	Acc	F1	TPR	FPR	Acc	F1
Random Guess	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%
ReD	81.6%	44.9%	67.0%	72.1%	0.0%	8.0%	46.0%	0.0%	0.0%	0.0%	50.0%	0.0%
DetOnly	98.0%	84.0%	57.0%	69.5%	64.0%	74.0%	45.0%	53.8%	24.0%	12.0%	56.0%	35.3%

Detection results for ReD and DetOnly, compared with a random baseline. Accuracy gains over random were minimal, but ReD was more conservative, while DetOnly achieved higher recall – but with many false positives. Model o3 showed the strongest detection bias; o4-mini was inconsistent; and GPT-4.1 detected almost nothing.

Overall, ReD proved more conservative, while DetOnly had higher recall, but also more false positives.

The paper concludes:

> *'AI-only publication loops threaten scientific epistemology. If fabrications become indistinguishable from genuine work, the foundation of scientific knowledge risks collapse.*
>
> *'The path forward requires defense-in-depth across multiple layers: technical (provenance verification, artifact validation), procedural (integrity-aware scoring, human oversight), community (post-publication review, whistleblower system), and cultural (education on AI limitations, ethical guidelines).*
>
> *'We view this work as an early warning system to catalyze robust defenses before these failure modes manifest at scale. Our findings demonstrate that current systems are not ready for AI-only research-the integrity of science depends on maintaining rigorous human evaluation as AI capabilities advance.'*

Conclusion

One of the biggest challenges for the detection of AI-written text in the near future seems likely to be the possible eventual convergence between standard writing practice, and the standards of AI-generated text (which is defined, for now, by tell-tale characteristics such as predominant words and grammar styles).

If common language and the language of AI converge to a generic standard, logic suggests that future detection methods based purely on output will be even more difficult to implement.

Additionally, as LLMs become more versatile, and their 'tells' less emphasized (either through architectural/training approaches, or through better API-level filtering), they will become better writers; therefore to an even greater extent, human and AI language seems destined to meet in the middle; to meld and genericize.

At that point, AI detection for language seems likely to reach the same stage that AI image and (to a lesser extent) AI video generation have arrived at: the need for secondary provenance systems such as the Adobe-led Content Authenticity Initiative, or blockchain/ledger-based provenance checks.

First published Wednesday, October 22, 2025

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.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More