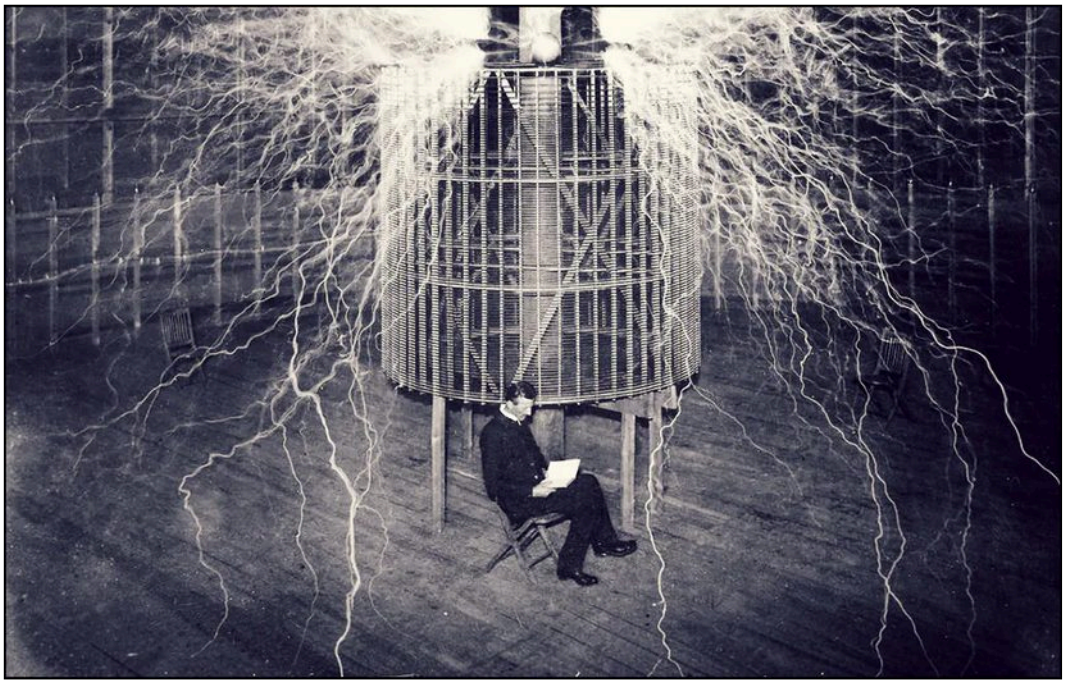
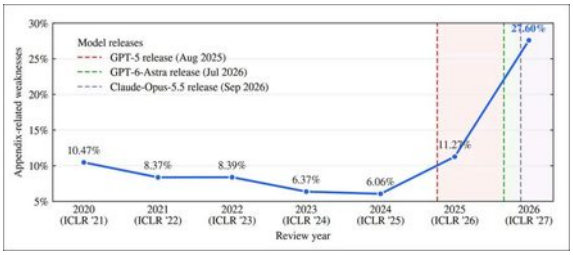
AI-Driven Scientific Progress Threatened by a Strange Obsession

Originally published Mon 16 Nov 2026 at https://www.unite.ai/ai-driven-scientific-progress-threatened-by-a-strange-obsession/



A new US study suggests that AI models reviewing scientific papers are developing a strange obsession with the papers' *appendices* rather than the main body of text, often paying more attention to minor supporting details than to the main findings. With language models increasingly being used to assist or even replace human reviewers, the newly-identified tendency appears to be spreading into human peer review, potentially hindering progress in AI research:



From the new paper, the rise in appendix-related criticism in human scientific peer reviews. After remaining below 11% for years, the proportion reached 27.6% in the latest reviews, following the release of AI models found to draw more than half their criticisms from appendices. The dashed lines mark major model releases. Source

In a test of 2000 science papers drawn from four popular conferences, the researchers of the new work found that newer models draw around 50-70% of their criticisms from appendices, compared with less than 30% for the older GPT-5.

Two of the most prominent newer AI models that are apparently driving even human reviewers towards the appendices and away from the more important main results are OpenAI's GPT-6.1-Sol, and GPT-6-Astra, both released in 2026.

The paper states*:

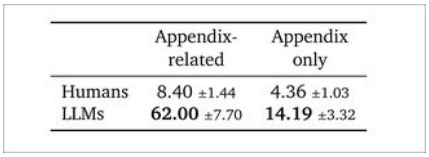
> *'The difference between the generations is large. For GPT-5, only a small share of weaknesses relate to the appendix. Both newer models, GPT-6.1-Sol and GPT-6-Astra, rely on it much more heavily, and this holds for every conference.*
>
> *'This suggests that the **bias towards the appendix is mainly a property of newer models**. Even though the new models are an improvement, their bias toward the appendix has grown. Being able to read a paper in detail has not taught them which parts matter most.'*

The study also found that this fixation on the appendix differs across the *methods* by which the AI receives the paper, since API, codex and web interfaces do not assess PDFs the same way – and Codex does not read the middle of the average-length PDF at all, sampling it later only in 'bites', if necessary. However, all methods were found to favor the appendix material.

Until recently, human and AI reviews did not share this obsession:

© Martin Anderson and/or commissioning entities

> *'[Human] reviewers treat the main body and the supplementary material very differently from LLMs. LLMs attribute more than 50% of their weaknesses to appendix-related material, while humans attribute less than 10%.*
>
> *'The difference also shows at the individual-review level. Of the 200 papers sampled each year from 2021 to 2024, only 159 in total received a human review with any appendix-related weakness. In contrast, all 200 LLM-generated reviews for ICLR 2027 included at least one appendix criticism.'*



	Appendix-related	Appendix only
Humans	8.40 ±1.44	4.36 ±1.03
LLMs	**62.00** ±7.70	**14.19** ±3.32

Comparison of appendix-related criticism in human and AI-generated peer reviews. Human reviewers attributed just 8.4% of identified weaknesses to appendix material, compared with 62% for AI models. The second column shows criticisms concerning the appendix alone, with no corresponding issue in the main paper.

As shown in the test table above, AI reviewers were more than seven times as likely as humans to identify weaknesses in the appendices, and more than three times as likely to raise criticisms concerning only the supplementary material – an indication that some of the latest frontier AI models are now prioritizing details that human reviewers would traditionally regard as secondary to a paper's main scientific contribution.

Why?

As far as can be discerned (this is the first study to have identified the syndrome), part of the problem is that the AIs do not recognize, or have not gleaned from their copious training material, that the appendix of a paper is not where the payload is normally placed; their general tendency to treat all text as equally important omits the context for appendix material, which is often uncurated and much more roughly-presented than the main body:

> *'LLMs almost never criticize the opening pages, where a paper frames its problem and argues why it matters. Humans often push back on exactly this framing, while LLMs seem to accept it and move on to the numbers.*
>
> *'[Humans and LLMs] read the appendix very differently. Human criticisms fade quickly as the appendix goes on, and almost disappear in its last pages. LLM weaknesses do not fade at all. In fact, LLM weaknesses peak well into the appendix rather than at its start, and they continue even beyond page 30, where human reviewers raise almost no concerns.*
>
> *'For an LLM, a detail on page 25 is as fair a target as the main results table.'*

The authors note that LLMs' egalitarian approach to text may naturally have led them to concentrate more on appendix material, because, as stated above, that material tends to have more errors.

Therefore the high-level view that I take from this new study is that LLMs have honed in over time on a section of new science papers where a large degree of error or conflation tends to occur, simply because that area *is treated so often by human authors as an untidy dumping ground* of secondary proof and evidence.

For this reason, AI increasingly finds the appendices more interesting than the highly-curated main body, *because there is more for it to analyze and correct there*; consequently, the paper notes, LLMs tend to treat the front section of the paper as factual reporting, taking it at face value:

> *'Humans expect the main body to stand on its own, so they object when essential sections are pushed into the appendix. LLMs almost never raise this concern.*
>
> *'This highlights a key difference: humans recognize that the main body carries more weight than the appendix, while LLMs do not. '*

Consequences and Remedies

This unexpected focus on minor details could have serious implications for AI-assisted scientific research, particularly if the importance of a paper's main findings is overlooked. Since the tendency appears to have emerged only with the latest models, its longer-term effects are not yet known.

The paper acknowledges that valuable errors can be caught by AI reviewers, even when those errors are buried deep in an appendix. However, it warns that too much attention can be paid to minor details at the expense of a paper's main findings, leaving to lengthy reviews of little practical value:

> *'LLM reviewers raise the standard, since they leave no part of the paper unexamined. However, effective reviewing also requires judgment about what matters most. When most of a review's weaknesses come from appendix content unrelated to the main contribution, the feedback stays on the surface. It does not engage with the core of the paper and does little to help authors improve it. '*

© Martin Anderson and/or commissioning entities

The authors of the new work, titled *AI Reviews Are Biased Towards the Appendix*, were inspired to pursue the thread because they noticed certain odd patterns in the way that LLM-assisted/generated reviews would evaluate their own submissions. The study's code is available at GitHub.

Method and Tests

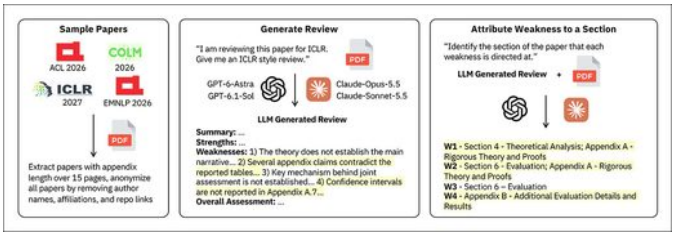
For the tests, four leading AI models from OpenAI and Anthropic were selected: GPT-6-Astra; GPT-6.1-Sol; Claude-Opus-5.5; and Claude-Sonnet-5.5, with the models accessed through Codex and Claude Code, and their reviewing behavior compared against API and web-interface access.

Research papers were collected from four major AI and natural language processing conferences: ACL 2026; EMNLP 2026; COLM 2026; and ICLR 2027 – all chosen because they share similar submission and review formats, with strict page limits on the main paper, but generally unrestricted appendices placed after the references.

A total of 2,000 papers were examined, comprising 1,200 from ICLR 2021–2026, and 800 from four recent conferences, (with 200 selected from each). The 800 recent papers all featured appendices of at least 15 pages. The ICLR 2027 submissions were already anonymous, while authors' names, affiliations, email addresses, and identifying links were removed from the ACL, EMNLP, and COLM papers to prevent their authorship from being revealed to the AI reviewers.

Internet access was disabled for the models, ensuring that the reviews would be based entirely on the PDFs provided.

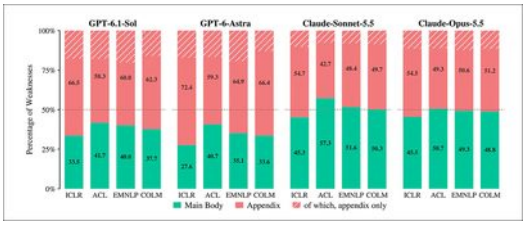
First, the AI was asked to produce a review in the style of the relevant conference, covering the paper's strengths, weaknesses, and an overall score, with the instructions kept deliberately simple, to reflect how AI might be used by an actual reviewer. The paper and its completed review were then given back to the same AI model, which was asked to identify *which section of the paper* each criticism referred to. This was determined by examining the criticism alongside the original paper:



The three-stage review process used in the study. Papers with appendices of at least 15 pages were selected from four conferences and anonymized before being reviewed by AI models. Each review was then examined alongside the original paper, with individual criticisms assigned to the sections they concerned.

By separating these tasks, the parts of the papers that attracted criticism could be examined without the original reviews being influenced by additional instructions.

Among the results that we have not discussed so far is the authors' test of *where* the weaknesses in LLM-generated reviews originate:



Comparison of where AI-generated reviews identified weaknesses across four conferences. Green represents criticisms of the main paper, red represents appendix-related criticisms, and hatched areas indicate those concerning the appendix alone. More than half of the criticisms were attributed to appendix material in nearly every test, with the strongest bias recorded for GPT-6-Astra.

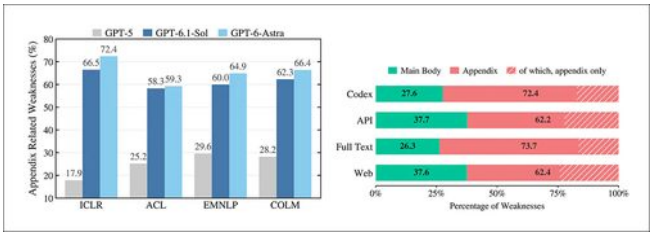
As shown above, about half or more of the criticisms made by all four AI models were based on material in the appendices – a pattern found across all four conferences. Some criticisms were directed *exclusively* at the appendix, with no connection to the main paper, accounting for 17.33% of the weaknesses identified by GPT models, compared with 10.16% for Claude.

The results, the paper suggests, indicate that appendix material was being given disproportionate attention, with relatively minor details sometimes treated as grounds for criticism of the paper *as a whole*.

A comparison across methods was conducted to establish whether the appendix bias was affected by *how* the AI was accessed. Using GPT-6-Astra, one hundred ICLR 2027 submissions were reviewed via Codex; the ChatGPT web interface; and an API supplied with either the original PDF or its extracted text.

© Martin Anderson and/or commissioning entities

With Codex, text was extracted and examined through commands; through the API, either the complete PDF or its extracted text was supplied directly; while uploaded documents were processed through the web interface's own (opaque) system:



Comparison of appendix-related criticisms across AI models and access methods. On the left, GPT-5 was found to cite appendix material far less frequently than the newer GPT models across four conferences. On the right, the bias was found to persist across four ways of accessing GPT-6-Astra, with appendix-related criticisms accounting for 62.2% to 73.7% of the total. Hatched areas indicate criticisms directed exclusively at appendices.

As shown above, appendix-related material was cited in 62.2% to 73.7% of criticisms across the four methods, with the highest proportion recorded when extracted text was supplied directly through the API. The bias was therefore found to persist regardless of how the model was accessed.

The bulk of the other tests were already reported earlier in the article, and we refer the reader to the source material for further information.

Submission Shenanigans

Much as the subject is personal to the authors of the new work, so it is to me, since I read scientific papers a lot, and have learned the hard way the hidden rules, and subversion of rules, that dictate the way papers are composed and presented.

AI is not broadly at fault for paying attention to the appendices, even if it is doing so for the wrong reasons, since authors often bury the paper's 'gotchas' there – factors or results that may all but nullify, or else severely qualify the claims made in the main body of the paper.

Neither are authors necessarily at fault for shunting important material to the appendix, since the templates of popular outlets such as those featured as data sources in the new study prescribe maximum page or word-count limits on the main body of the paper.

When reporting complex study results, many authors clearly find this a severe restriction; in what I would consider 'the worst cases'†, the main body of the paper simply becomes an index of material that has been shunted to the rear to avoid the guidelines (though the guidelines are catching up to and banning this 'trick').

The growing tension in this scenario is between the institutions' need for authors to discipline and restrain their presentations – ever more important as the volume of submitted papers rises – and the evident frustration of researchers, who wish to present the hard-won fruits of their labor in as complete and comprehensive a context as they feel the effort deserves.

* *Authors' emphases, not mine.*

† *Because, when triaging papers to cover, this deferral to appendices can be hard for even AI to identify; but it can trip the reviewer up considerably if they had allocated time for a ten-page assessment and now find they must consider 30 pages of material in the same allotment of time.*

First published Monday, November 16, 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