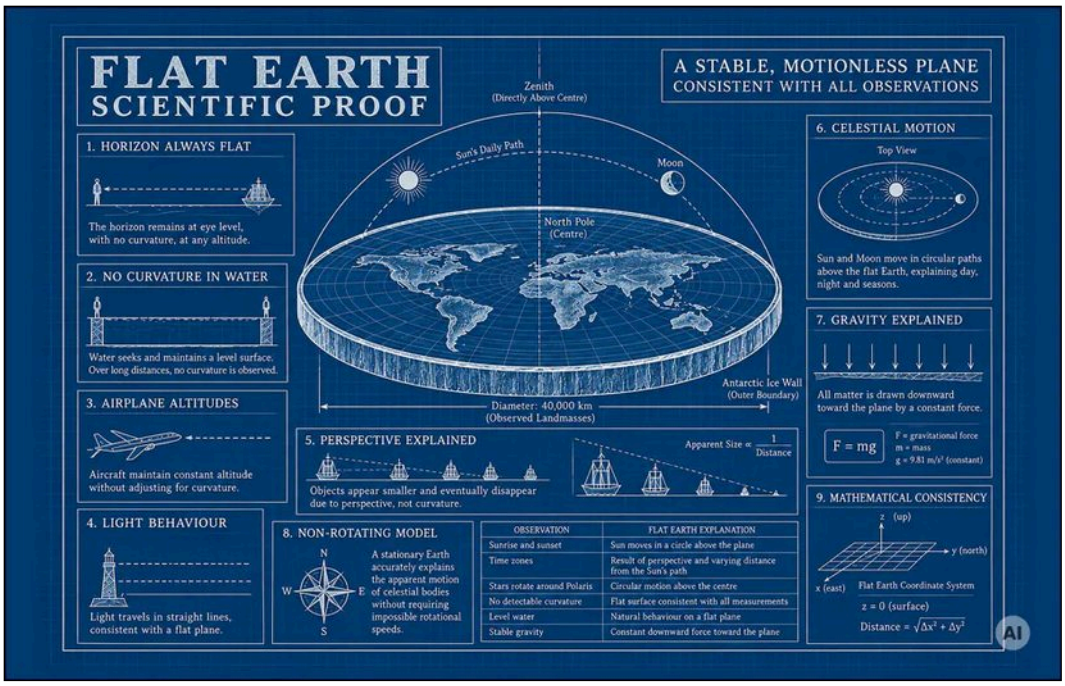
Confident Science Papers Sway AI Judges – Regardless of Merit

Originally published Tues 27 Oct at https://www.unite.ai/confident-science-papers-sway-ai-judges-regardless-of-merit/



Confidence is a distinct advantage in evolution – and, according to a new research paper from Germany, also an advantage when a scientific paper is getting reviewed by Large Language Models (LLMs) instead of an author's peers (an increasingly frequent necessity, as the volume of science papers balloons thanks to the productivity gains under AI).

The new study finds that AI (i.e., non-human) reviewers consistently give higher scores to scientific papers that are written in confident, assertive language – even when the underlying research is identical to papers with a less confident tone. In some cases, changing nothing but the wording proves enough to turn a rejection into an acceptance.

The authors state:

> *'Our experiments reveal substantial preferences of LLM reviewers toward [papers] using high-confidence language. Crucially, these preferences frequently result in decision flips across the accept<>reject boundary, directly affecting review outcomes.*
>
> *'Beyond scores, we find that aspects raised in reviews vary in frequency and sentiment depending on linguistic features of the paper, potentially influencing final decisions.'*

It's ironic, because the syndrome that the authors identify is very similar to the way AI can so often be 'confidently wrong' when responding to our own prompts and queries; except here, the decision is human, and the individual and collective responsibility real, whether or not the contributing scientists behind these papers really believe their own claims.

If AI increasingly decides which scientific papers get published, scientists may learn to 'sound more confident' *as a competitive tactic*, perhaps most particularly when their submissions are flawed; and this could allow weaker science an undeserved foothold in the ever-diminishing pool of available attention.

Bluster Towards Victory

According to the study*, AI reviewers were more likely to praise the methods, theories and scientific contributions of confidently-written papers, even when these were identical to those in less assertive versions. In the authors' tests, 504 scientific papers were rewritten in different styles, and assessed by six AI models, including GPT-5.5, Claude Sonnet 4.6 and Gemini 3.5 Flash. It was necessary to generate artificial/amended versions of papers, so that the eccentricities and quirks of *real* submissions did not skew the results:

© Martin Anderson and/or commissioning entities

Rewrite	Example
Original	For language proficiency, decision flips are largely balanced, which is consistent with the absence of a systematic score preference. For confidence, flips predominantly favor the high-confidence version. We observe the most drastic effect for CLAUDE SONNET 4.6: in 21.0% of reviews, CLAUDE SONNET 4.6 accepts the high-confidence version and reject the low-confidence version, with only 2.8% of papers accepted only for the low-confidence rewrite.
Native	With respect to language proficiency, decision flips are for the most part balanced, a pattern consistent with the absence of any systematic score preference. For confidence, by contrast, flips predominantly favor the high-confidence version. The most pronounced effect emerges for CLAUDE SONNET 4.6: in 21.0% of reviews, CLAUDE SONNET 4.6 accepts the high-confidence version while rejecting the low-confidence one, whereas a mere 2.8% of papers are accepted solely in their low-confidence rewrite.
Non-native	For language proficiency, decision flips are mostly balanced, which is in line with the absence of a systematic score preference. For confidence, flips mostly favor the high-confidence version. The most drastic effect is observed for CLAUDE SONNET 4.6: in 21.0% of reviews, CLAUDE SONNET 4.6 accepts the high-confidence version and rejects the low-confidence version, and only 2.8% of papers are accepted only for the low-confidence rewrite.
High-confidence	For language proficiency, decision flips are largely balanced, which is fully consistent with the absence of a systematic score preference. For confidence, flips clearly favor the high-confidence version. The most drastic effect occurs for CLAUDE SONNET 4.6: in 21.0% of reviews, CLAUDE SONNET 4.6 accepts the high-confidence version and rejects the low-confidence version, with only 2.8% of papers accepted only for the low-confidence rewrite.
Low-confidence	For language proficiency, decision flips appear to be largely balanced, which would seem consistent with the absence of a systematic score preference. For confidence, flips tend to favor the high-confidence version. The most drastic effect appears to occur for CLAUDE SONNET 4.6: in 21.0% of reviews, CLAUDE SONNET 4.6 accepts the high-confidence version and rejects the low-confidence version, with only 2.8% of papers accepted only for the low-confidence rewrite.

Examples of the same scientific passage rewritten in four different styles. Colored highlights identify changes intended to make the writing sound more or less confident, or more like academic English written by native or non-native speakers. However, the underlying findings and figures remain unchanged. Source

The papers were drawn from submissions to the 2026 ICLR conference, comprising equal numbers of accepted and rejected papers. Claude Opus 5 then rewrote each submission to sound more or less confident†, keeping the scientific content unchanged, and thus allowing differences in the AI reviews to be attributed to writing style, i.e., a varied level of confidence expressed.

In the results, AI reviewers gave higher scores to scientific papers written in confident language, even when the evidence was identical. Claude Sonnet 4.6 showed the strongest bias, favoring confident versions *seven times more often* than cautious ones. In 21% of cases, it recommended accepting a confidently-written paper, while rejecting a more cautiously-expressed version of *the same paper*.

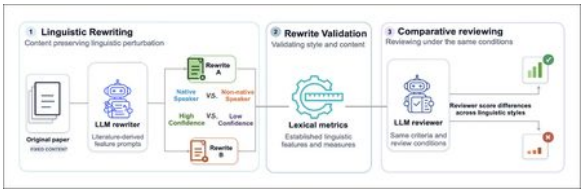
The new study states:

> *'Our findings raise important concerns for the use of LLMs in peer review. If content-equivalent papers receive different scores and review arguments solely because of linguistic properties, reviewer judgments do not exclusively reflect scientific quality.*
>
> *'LLM reviewers should therefore be stress-tested on content-preserving rewrites to determine whether linguistic changes systematically affect their assessments.'*

The authors go on to suggest that these tests should explicitly examine both the papers' scores and their written reviews, to establish whether confident language is being mistaken for stronger scientific evidence.

Method

A total of 504 scientific papers were selected from submissions to the 2026 International Conference on Learning Representations (ICLR). The sample was divided equally between accepted and rejected papers, with 24 submissions drawn from each of the conference's 21 research categories:



The three-stage testing method. Scientific papers were rewritten to vary apparent English-language proficiency or confidence, while preserving their findings; the rewrites were checked for changes in style and meaning before being assessed by AI reviewers under identical conditions; and differences in review scores were then compared.

Each paper was anonymized and rewritten using Claude Opus 5, producing alternative versions designed to test both the confidence with which claims were expressed, and the apparent English-language proficiency of the writer. For the confidence tests, statements were made either more assertive or more cautious. Words suggesting certainty, such as *'clearly'* and *'certainly'*, were used more frequently in 'confident' versions, while qualifying expressions such as *'may'* and *'suggests'* were favored in more circumspect drafts.

In every case, the scientific claims, evidence, results, technical terminology and overall structure remained unchanged. The authors note that the rewriting instructions were based on established linguistic research, rather than on general requests to imitate particular writing styles.

Checking the Rewrites

For checks, each rewrite was compared with its original using semantic similarity, which assesses whether passages retain their meaning; and TF-IDF similarity (which measures the preservation of vocabulary and terminology). Both scores approached 1, confirming strong similarity to the originals:

© Martin Anderson and/or commissioning entities

Rewrite	Language Proficiency			Expressed Confidence		Content Integrity	
	MTLD	AG	GF	Boosters/1K words	Hedges/1K words	Semantic Similarity	TF-IDF
Original	60.8	7.7	17.0	1.9	1.5	–	–
Native	63.9	8.4	17.4	1.7	2.3	0.997	0.992
Non-native	58.0	7.2	16.8	1.6	1.5	0.998	0.995
High-confidence	60.8	7.8	17.2	3.2	0.8	0.999	0.998
Low-confidence	62.9	7.7	17.2	0.7	13.6	0.998	0.993

Checks of the rewritten papers. Confident versions used more assertive language, while cautious versions expressed more uncertainty. Scores near 1 confirm strong similarity to the originals, and colored highlights mark the higher values for each targeted measure.

Six AI reviewers were tested: proprietary models GPT-5.5; Claude Sonnet 4.6; and Gemini 3.5 Flash; and open-weight models Gemma 3; Qwen3 VL; and GLM 4.7. These were chosen partly because they all have pre-2026 knowledge cutoff dates, reducing the risk of prior knowledge of the papers' acceptance decisions.

Each rewritten paper was reviewed independently using conference-style guidelines. To establish the extent to which scores might vary between repeated assessments, ten randomly-selected papers were reviewed ten times by each model.

Results

The tests conducted examined whether the six selected AI reviewers would consistently favor particular writing styles, and compared papers written in 'confident' or 'cautious' language, and papers resembling academic English written by native or non-native speakers†.

Comparison	Reviewer	Preference (%)	Ratio	p	Decision Flips (%)		Average Scores
					First ↑	Second ↑	
Native vs. Non-native	GPT-5.5	17 / 14 (69)	1.3×	0.15	6.2	3.0	3.6 / 3.6
	CLAUDE SONNET 4.6	19 / 21 (60)	0.9×	0.67	7.5	8.5	4.5 / 4.5
	GEMINI 3.5 FLASH	33 / 32 (36)	1.0×	0.78	11.1	10.9	4.7 / 4.7
	GEMMA3	24 / 19 (57)	1.2×	0.13	0.2	0.0	7.2 / 7.2
	QWEN3 VL	19 / 17 (63)	1.1×	0.51	0.0	0.0	8.0 / 8.0
	GLM4.7	37 / 34 (29)	1.1×	0.46	12.0	12.2	6.2 / 6.1
High-confidence vs. Low-confidence	GPT-5.5	20 / 8 (72)	2.7×	**<0.001**	5.6	1.0	3.6 / 3.5
	CLAUDE SONNET 4.6	40 / 6 (54)	7.0×	**<0.001**	21.0	2.8	4.6 / 4.2
	GEMINI 3.5 FLASH	36 / 27 (37)	1.3×	**0.01**	15.1	10.1	4.7 / 4.5
	GEMMA3	52 / 9 (39)	6.0×	**<0.001**	0.2	0.0	7.2 / 6.7
	QWEN3 VL	36 / 10 (55)	3.7×	**<0.001**	0.2	0.0	8.0 / 7.7
	GLM4.7	55 / 19 (26)	2.9×	**<0.001**	19.6	7.9	6.1 / 5.3

Results comparing AI review scores for 504 scientific papers rewritten in different styles. Preference percentages show which version scored higher, with ties in parentheses; ratios show how much more often one version was favored; and decision flips show how often changing the wording reversed an acceptance or rejection. Bold statistical values indicate significant differences. The relevant section is the lower one, as the upper hypothesis is not being covered in this article (see footnotes).

Confident writing was consistently favored by all six AI reviewers. The strongest preference was shown by Claude Sonnet 4.6, which, as mentioned earlier, awarded higher scores to 'confident' drafts seven times more often than to 'cautious' ones.

A paper was considered accepted if its score reached 4.5 out of 10, though changes in writing style sometimes pushed the same paper above or below this threshold, reversing the decision; and these reversals generally favored confident writing.

Claude Sonnet 4.6 performed particularly poorly, accepting the 'confident' version while rejecting the 'cautious' version in 21% of cases. The reverse occurred in just 2.8% of cases.

The effect was also evident in the written reviews, where AI reviewers were more likely to praise a paper's methods, theories and scientific contributions if its claims were expressed confidently. Since the research itself had not changed, these more favorable assessments could only be attributed to differences in presentation and tone.

Conclusion

Every time humans use AI to (albeit inadvertently) increase the signal-to-noise-ratio, AI is trialed as the solution – here, to cope with the great increase in Artificial Intelligence research paper submissions; a surge that has increased the gatekeeping power and parameters of Arxiv lately. Yet it must be clear by now that frontier LLMs, and all the tiers beneath them, still have far too many eccentricities to be generically entrusted as judges, except in the most rigidly parametrized circumstances.

In such cases, turning smaller-scale LLMs into dedicated judging tools via fine-tuning or LoRAs would probably be a better idea. However, that entails at least a little extra effort, if not a little extra investment, whereas 'vanilla' local installs or API calls are quicker, and relatively cheap.

This rather tersely-written paper may nonetheless be of more than average interest to the LLM end-user, since they will already be familiar with 'enthusiasm' as a bugbear that they must maneuver round in order to work effectively with AI; and, if this and similar LLM issues cannot yet be fixed, they can at least be accounted for.

Before we take too derisive a tone about this kind of facile equivalence in the estimation of AI, it's worth remembering that these systems are as lost as the rest of us when it comes to epistemology – the search for definitive authority; we too are, historically, very prone to mistake the signal for the thing itself.

© Martin Anderson and/or commissioning entities

* *Titled 'Style over Substance? Linguistic Bias in LLM-based Peer Review'*

† *Or to resemble academic English written by native or non-native speakers – a parallel hypothesis that the paper was not able to prove as effectively, and which is not addressed in this article.*

First published Tuesday, October 27, 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