

AI Article Reviews Dominate Human Peer Reviews at Their Best: Evidence and Policy Choices*

Gary King[†]

Steven Worthington[‡]

October 28, 2026

Abstract

We show that AI-generated article reviews are now better, by every measure we could construct, than reviews written by well-compensated human experts and judged to be of unusually high quality by journal editors and authors. We then discuss how this result may inform journal policy choices, with concerning downsides as well as potential paths to increase publication speed and quality; reduce adversarial interactions; improve the experience of editors, reviewers, authors, and readers; and advance science.

*Our thanks to Lee Bright, Mitchell Duneier, Felix Elwert, Chris Winship, and various AI models, for helpful comments. The most recent version of this paper and its Supplementary Appendix is available at GaryKing.org/peerless.

[†]Albert J. Weatherhead III University Professor, and director of the Institute for Quantitative Social Science, Harvard University; GaryKing.org, King@Harvard.edu,  0000-0002-5327-7631

[‡]Director of Data Science, Institute for Quantitative Social Science; StevenWorthington.org; SWorthington@IQ.Harvard.edu, Harvard University.  0000-0001-9550-5797

1 Introduction

Human peer review is the default quality-control standard for scholarly publishing, and yet problems with this system are deep, common, and well known (Aczel, Barwich, et al., 2025), including self-serving reviewer conduct, weak incentives and lack of expertise for some reviewers, and insufficient scrutiny; low inter-reviewer agreement (Bornmann, Mutz, and Daniel, 2010; Mutz, Bornmann, and Daniel, 2012); bias due to author characteristics, topics, and methods, with opacity and little accountability (Tennant and Ross-Hellauer, 2020); failure to detect major errors and little overall evidence of improved quality (Gasparyan and Kitas, 2012); frustrating delays (Farber, 2025); large costs in reviewer time (Aczel, Szaszi, and Holcombe, 2021); reviewer recruitment difficulties (Wright, 2024); unhelpful feedback (Chong and Lin, 2024); and resistance to heterodox findings (Armstrong, 1997).

These problems with the human peer review process have motivated scholars to evaluate large language models (now often called AI) as a replacement. But AI leads to other worries, including hallucinated claims, fabricated sources, and confidently stated errors (Ye et al., 2024); failure to recognize or reward novelty (Shin et al., 2025); biases learned from human training data unfairly favoring authors based on institution, seniority, or fame (Macharla Vasu et al., 2026; Hosseini and Horbach, 2023); excessive leniency (Russo Latona et al., 2024); vulnerability to manipulation (Lin, 2025; Ye et al., 2024); breaches of confidentiality (Naddaf, 2025); and unclear accountability for errors (Hosseini and Horbach, 2023; Porsdam Mann et al., 2025). The farthest scholars have gone is to conclude that AI should supplement rather than replace human peer reviews (Liang et al., 2024; Farber, 2025; Doslaliuk et al., 2025), including the largest deployment to date, at the AAAI-26 conference, which added one clearly labeled AI review to the human reviews of each of the 22,977 papers sent for full review (Biswas et al., 2026). Two recent comparisons of AI and human reviews of the same journal manuscripts point in opposite directions. For 119 articles in *BMJ Open*, human reviewers' reports raised more detailed and varied substantive points than ChatGPT's (Rajakumar et al., 2026). For 40 manuscripts at a cardiology journal, two editors who did not know which reviews were

AI-written rated reviews by Claude, ChatGPT, and Gemini higher than the human reviews on five of seven quality criteria (Zancanaro et al., 2026).

In this paper, we conduct head-to-head experiments using an unusually high quality set of human reviews, and show that AI now dominates human peer reviews at their best — according to blind judgments by other AI models in every area and by (human) former editors of major scholarly journals. Given prior research on earlier AI models, we had expected AI to win some in some areas, some of the time, but we found that the latest AI wins in them all, by every criterion we measured, by wide margins, with no fabricated sources, and with about a third the rate of incorrect claims of the human reviews. We did not measure the degree to which subsequent revisions improved the paper or the collective consequences of switching to AI reviewing. We describe our research design in Section 2, statistical results in Section 3, and robustness checks in Section 4. We discuss the remaining threats to different inferences from our analysis along with potential implications of our findings for journal policy in Section 5.

2 Research Design

We now discuss the papers reviewed, the AI and human paper reviewers, and how we used AI and human judges to separately choose which review of each paper is better (see Supplementary Appendix A for details).

Paper Submissions and Human Reviews For this study, we used papers submitted to *rapidPeer*, a service developed by Harvard University’s Institute for Quantitative Social Science (and then expanded to several other institutions) that gives university faculty high quality double-blind reviews of their paper drafts prior to formal journal submission. The service asks human reviewers to read the entire paper and provide constructive suggestions throughout (as distinct from behaviors we sometimes see from journal reviewers, such as searching for a fatal flaw on p.3 so they can stop reading, or making it more about their own work than appropriate). Reviewers are told not to focus on whether the paper is suitable for a particular journal or meets a specific standard (although publication sugges-

tions are welcomed) and instead asked to be comprehensive, to write at least 1.5-2 pages, and to be helpful to the author when revising. The instructions suggest a wide range of ways for reviewers to help (see Supplementary Appendix B for the exact language).

Reviewers are carefully chosen by experts in the field from institutions other than their own and incited to provide highly constructive reviews: They are paid (well) for their time and told they are likely to have another opportunity to be paid (or to submit a paper to the process themselves) if they do a good job. All manuscripts are deidentified prior to the review process. Whether judged by us, PhDs on the *rapidPeer* staff, or the paper authors (who fill out evaluations upon receipt), these reviews are of exceptionally high quality: 88% of the 74 authors who answered a survey a week after delivery rated the review they received as high or very high quality, and the median author rated their willingness to recommend the *rapidPeer* service to other faculty at 10 out of 10 (Table 2). We also informally consulted former editors from major journals, all of whom told us that the reviews are among the best they have seen in their times as editors. We know of no collection of human reviews written under better conditions, and none that authors or editors rated more highly.

The corpus includes every eligible submission to *RapidPeer*, including 124 scholarly paper submissions and 134 human reviews (10 submissions received two independent reviews) collected between 2019 and 2026 and drawn from a wide range of research, including statistical analyses of empirical data, formal mathematical modeling, and purely qualitative studies. Near as we can tell, all reviews were written by humans: many predated ChatGPT; neither the quality nor style changed over time; and reviewers were compensated and responsible for their contribution.

AI Reviews For each submission, we generated a single AI review with Claude Opus 4.8 (Anthropic, 2026) at its maximum reasoning setting, under two conditions. In the *matched* condition the model received the anonymized manuscript (including figures and tables) together with the verbatim *rapidPeer* reviewer guidelines (Supplementary Appendix B) as its instructions and no external tools. Like the human reviewers, it worked from the manuscript and guidelines, with two differences: human reviewers could also consult

supplementary files, and the AI’s training data may include work published after the submission, the effects of which we test for in Section 4.11. In the *tuned* condition, the AI model received the same manuscript but with a purpose-built agent prompt tuned to the AI along with web retrieval abilities (Supplementary Appendix C). The tuned prompt was written before any judging, was never tuned against judges’ scores, and does not contain the judging rubric. Every tuned review invoked web retrieval (mean 12.4 searches per review). Each review was produced in an isolated call — one review per submission per condition.

AI and Human Judges A model from a different family, GPT-5.5 (OpenAI, 2026), served as the AI judge, so that generator and judge share no lineage. AI judges may favor AI reviews for two reasons: AI models can prefer text written by other AI models (Laurito et al., 2025), and aggregate quality scores can overstate the quality of AI reviews that are fluent but generic or poorly grounded in the manuscript (Fichtl et al., 2026). We test for the first by rewriting the human reviews in the AI’s style and for the second by checking each review’s claims against the manuscript, and we also had human editors judge a sample of pairs (Section 4). Before comparing the two reviews, we had the AI decide which review topics were relevant to each paper (e.g., so a paper with no figures was not judged on “figures and tables”). We locked in that list of applicable topics once per paper, based only on the paper. Then we had the AI judge evaluate both reviews on only that set of topics. The AI judge evaluated the reviews under three instruments — an overall ranking of the two reviews, made from the reviews alone; a head-to-head ranking of the two reviews in each of seven review areas; and an independent individual rating score from 0 to 10 for each of the areas, the last two made with the manuscript’s text in hand (Supplementary Appendix D). The overall ranking judged each review as a whole, so it reflects how a review connects a paper’s framing, results, and figures together. Every comparison was run in both presentation orders, and a win was credited only when the same review was preferred under both orders; the few order-inconsistent pairs, and pairs in which neither review addressed an area, were treated as undecided. (We gave reviews silent on a relevant area a rating of 0; in head-to-head rankings, judges recorded

no winner for an area only when neither review usefully engaged it.) Per-area win rates are AI wins divided by decided pairs, and rating values are means, each computed only over the submissions for which the area is relevant, so denominators vary across areas. The seven review judgment areas — theory and idea, methods, weaknesses, framing, communication, figures and tables, and sources — were taken from the explicit *rapidPeer* reviewer guidelines rather than imposed externally (Supplementary Appendix E).

Finally, we enlisted three former editors of leading scholarly journals (in similar areas as the papers) to independently judge 10 randomly selected and blinded review pairs, one AI and one human but ordered randomly, and under the same overall ranking instructions given to the AI judge. The editors were unaware they were comparing AI to human reviews (until we debriefed them afterward). We did not attempt to ask our human editors to evaluate each review area separately (although debriefings with each editor revealed no differences in any area from the overall results). We collected this information via a web interface we designed for ease-of-use, with reviews for each pair presented side by side with independent scroll bars. On the AI judge’s ratings, the ten papers resembled the rest of the corpus: human reviews averaged 6.03 on the 0–10 rating scale in the editor sample and 6.02 in the other 114 papers, and the AI-human gaps were 2.67 and 2.64 points.

3 Results

We now summarize in Figure 1 an analysis of the 134 pairs of AI and human reviews, each pair of the same manuscript. The figure covers pairwise choices (Panel A), ratings of each item independently (Panel B), overall judgments (top), and AI ratings in seven review areas (bottom).

The results are unambiguous: No matter how we look at the data, AI wins. It wins in 100% of the pairs overall when judged by another AI model (132 of the 132 decided pairs)¹ and 87% overall when judged by human journal editors (Panel A). It wins in 92%

¹To simplify the presentation, the main text omits uncertainty estimates. Our 124 papers are a census of eligible *rapidPeer* submissions; if treated as a random sample from the population of similar papers, the exact (Clopper–Pearson) 95% confidence interval for the population proportion can be approximated from the reported proportion, p , and the number of independent units, n . When $p = 1$ (or 0), the interval extends in only one direction, and its lower end is at least $1 - 3.7/n$ (or its upper end at most $3.7/n$); for example, $p = 1$ with $n = 130$ gives $[0.972, 1]$. For other values, the interval is approximately $p \pm 2\sqrt{p(1-p)/n}$,

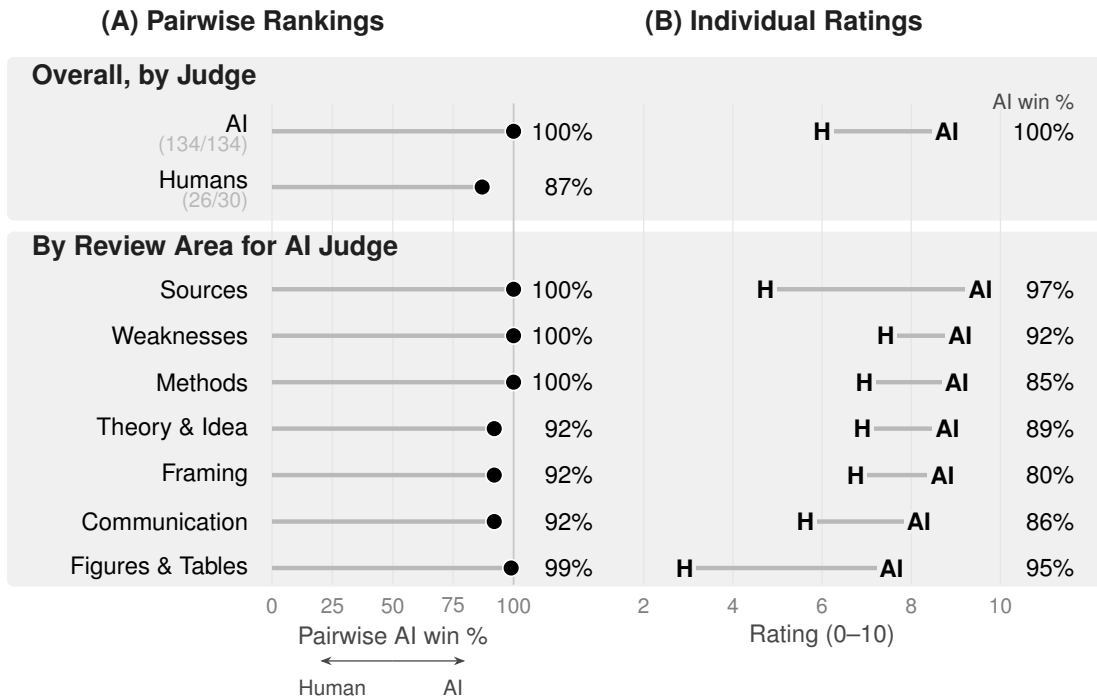


Figure 1: **AI reviews dominate human reviews.** The figure summarizes evaluations of 134 pairs of human and AI reviews of the same manuscript. The top reports overall evaluations by AI and by three (human) editors of major journals; the bottom summarizes results for seven separate review characteristics. Panel A gives pairwise rankings, the percentage of pairs where the AI review is ranked higher than the human review. Panel B gives mean ratings for both the human (**H**) and AI (**AI**) reviews.

of AI-Human review pairs for three of the review areas (theory and idea, framing, and communication), 99% for one (figures and tables), and 100% in three (sources, weaknesses, and methods).

Although human coding has long been known to be notoriously error-prone, we found that humans agreed as much or more with AI’s judgment than they did with other humans: human editors agreed with other human editors 80% of the time but on average 87% of the time with the AI’s judgment.

truncated at 0 and 1, or about ± 0.09 for p near 0.5 and ± 0.05 near 0.9 when n is 130; when n is about 100 or more, each end of the exact interval lies within 0.025 of this approximation, with the larger difference mostly at the end nearer 0.5, and for smaller n the differences can be larger. These intervals require the n units to be independent. Review pairs nearly are: the 10 papers with two human reviews contribute two pairs that share one AI review, which changes the intervals by less than 0.005. Units that share a judge, a review, or a pair are not independent, including the editors’ judgments (three editors judged the same 10 pairs), claims within a review, the replication’s comparisons of each human review with four AI reviews, and results pooled across review areas; intervals for those proportions would be wider than these formulas give.

4 Robustness Checks

We ran fourteen robustness checks, each testing whether the AI’s advantage in Section 3 comes from a feature of our research design rather than from better reviews. None change the conclusion.

4.1 Review Length

The AI does not win merely by writing at greater length. When we restricted the overall ranking to pairs whose AI-to-human word count ratio fell in $[0.7, 1.4]$, AI was preferred in all 44 of these length-matched pairs (100%), matching the full-corpus result.

4.2 Review Area Coverage

AI does not win only by covering more review areas: comparing ratings from pairs where both reviews addressed an area, the AI review had a higher mean rating in all seven areas (Table 1, panel A). We counted a review as addressing an area when it received an individual rating above zero there, the rubric’s indication that the review addressed the area. For an evaluation of figures and tables, the AI advantage narrowed from 4.54 points among all relevant pairs to 3.40 among the 90 pairs where both reviews addressed the area. Broader coverage therefore does not fully account for the observed rating advantage. Figure 2 shows the full distribution of individual ratings in each area, including the ratings of 0 given to reviews that did not address it.

4.3 The Strongest Human Reviews

AI wins even after (intentionally) biasing the sample in favor of the human reviewers via post hoc selection of the top human reviews: among the 28 pairs whose human reviews were most able to compete with the AI, the human review never outscored the AI. We selected these pairs by taking the top 40% of the 134 human reviews by the AI judge’s mean individual rating, and among those the 27 submissions (28 pairs) whose human review scored closest to the AI. This subset is chosen to favor the human on both quality and competitiveness, and it reuses the judge’s full-corpus scores with no additional judg-

Table 1: **Area comparisons after restricting coverage and retaining unresolved pairs.** Panel A compares mean AI–human rating gaps (0–10) over all relevant pairs and over pairs where both reviews addressed the area (both ratings above zero). Panel B counts AI wins over all relevant pairs, including unresolved cases. For each review area, AI results use the tuned (web-enabled) condition.

<i>A. Rating gaps when both reviews address the area</i>			
Review area	All relevant pairs	Both address the area	Pairs where both address it
Theory & idea	1.84	1.78	129
Methods	2.02	1.97	132
Weaknesses	1.62	1.62	134
Framing	1.92	1.87	129
Communication	2.48	2.43	129
Figures & tables	4.54	3.40	90
Additional sources	4.84	4.42	91
<i>B. Ranking wins with unresolved pairs retained</i>			
Review area	AI wins / all relevant pairs	Unresolved pairs	AI win %
Theory & idea	116/130	4	89.2
Methods	132/133	1	99.2
Weaknesses	134/134	0	100.0
Framing	109/130	12	83.8
Communication	111/130	10	85.4
Figures & tables	120/132	11	90.9
Additional sources	99/99	0	100.0

Panel A gives each review pair equal weight, so an AI review is counted twice for a paper with two human reviews; the separate AI means in Table 7 count each AI review once. Panel B’s unresolved pairs are neither AI nor human wins.

ing. In the tuned condition, even the single highest-rated human review scored 0.7 points below the AI, and the average human review in the subset scored 1.6 points below it on the 0–10 individual-rating scale; the AI also scored higher in every review area (by 1.0 to 3.6 points).

4.4 Authors’ Ratings of the Human Reviews

The human reviews are not a weak baseline: the authors who received them rated them highly. Beyond the survey results given in Section 2, 60% of the 74 responding authors rated their review’s quality very high, and the median author rated its impact 8 out of 10 (Table 2). Because *rapidPeer* delivered the human reviews to authors — the AI reviews were generated only for this study and never delivered — the AI’s advantage is measured

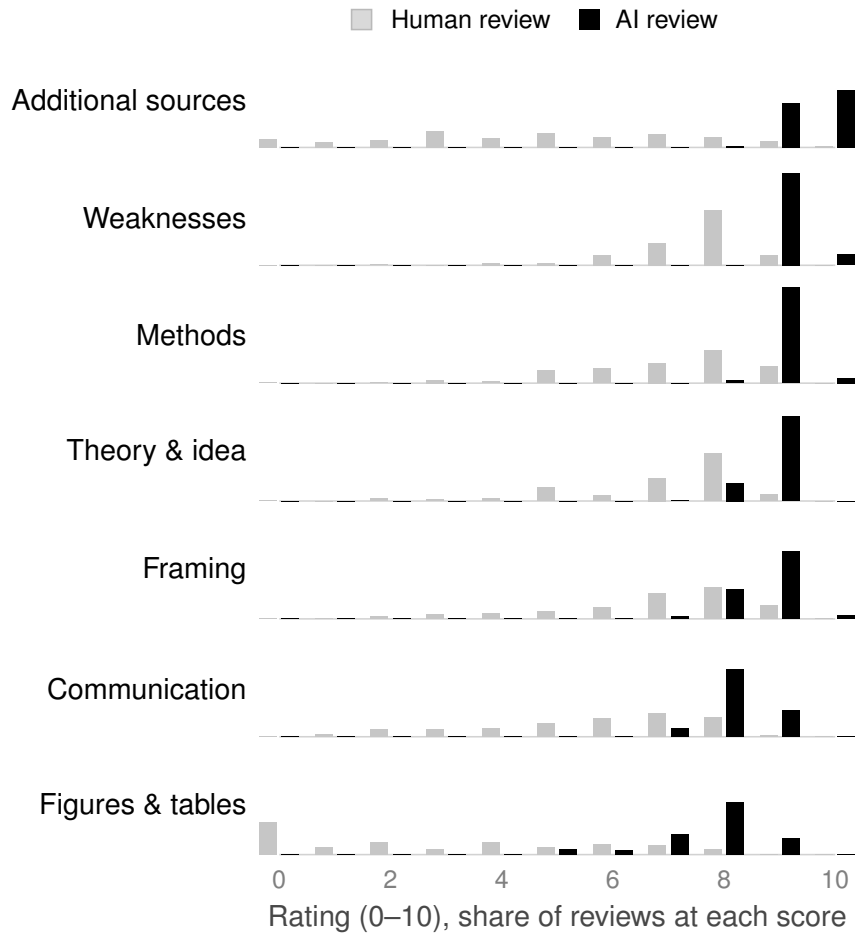


Figure 2: **Distribution of individual ratings by review area.** For each of the seven review areas, the bars show the share of reviews receiving each individual rating from 0 to 10, for the human reviews (light) and the AI reviews (dark). The AI reviews were written by Claude Opus 4.8 with access to web-retrieval, and every review — human and AI — was scored 0–10 by a blinded AI judge from a different model family, GPT-5.5, on only the review areas judged relevant to each paper. The distributions cover the 134 human and 124 AI reviews, with fewer in any area that is not relevant to a given paper.

against reviews their recipients valued highly.

4.5 Human Editors as Judges

AI judges can prefer text written by AI models (Section 2), so we checked the AI judge against human judges, and the human judges also preferred the AI reviews. The three former journal editors described in Section 2 preferred the AI review in 26 of 30 judgments (87%), and each independently favored the AI in a large majority of pairs (8, 8, and

Table 2: **Author evaluation of the delivered human reviews.** Survey one week after delivery, 74 respondents. The quality item was answered by all 74; the 0–10 impact and recommendation items by 57.

Measure	Result
Review quality high or very high	88% (65 / 74)
— very high quality	60% (44 / 74)
Impact of the review (0–10)	median 8 (mean 7.3)
Would recommend to a colleague (0–10)	median 10 (mean 9.4)

10 out of 10). We did not ask the editors to also review the associated manuscript. The comparison in Section 2 of the ten-paper editor sample with the remaining 114 papers averages the AI judge’s individual ratings over relevant review areas and gives each paper equal weight, averaging its human reviews when there are two; on these measures, the editor sample had similar human-review quality and a similar AI advantage.

4.6 Judge Access to the Manuscript

The AI judge described above made its overall ranking from the reviews alone, so we repeated the experiment, giving the judge the full manuscript and instructions to check each review’s claims against it, and the AI review was still preferred in 133 of 134 matched pairs and 132 of 134 tuned pairs. In the other three, the judge’s choice reversed with presentation order, so the human review was never preferred in both orders; counting these pairs as undecided, AI won 133 of 133 decided matched pairs and 132 of 132 decided tuned pairs.

Table 3 reports the overall ranking with and without the manuscript, for both matched and tuned reviews; each of the four comparisons covered all 134 review pairs in both presentation orders. Without manuscript access, the judge used the reviews-only prompt in Supplementary Appendix D.2. With it, the judge received all PDF pages as images and extracted text, with the prompt in Supplementary Appendix D.5. All four comparisons used the same GPT-5.5 judge and maximum effort setting, in a separate context for each judgment and with review authorship concealed, and the judge itself used no external tools.

Table 3: **Overall comparisons with and without manuscript access.** Each row covers 134 review pairs, each judged in both presentation orders. A win requires the same review to be preferred in both orders. Undecided pairs favor different reviews when order is reversed. All four comparisons used the same GPT-5.5 judge and maximum effort setting. Matched is the AI reviewer given no external tools. Tuned is the web-enabled review condition.

Condition	Manuscript	AI wins	Human wins	Undecided	AI wins / decided pairs
Matched	No	134	0	0	134/134
Tuned	No	132	0	2	132/132
Matched	Yes	133	0	1	133/133
Tuned	Yes	132	0	2	132/132

4.7 Writing Style

AI vs human writing style (according to formatting or writing register) had no effect, which we showed by rewriting human reviews in AI style and finding no improvement in overall judgments. An AI judge could favor AI-styled text in two ways: through surface formatting or through writing register. Since our GPT-5.5 judge saw both reviews with all formatting stripped, surface formatting differences are unlikely to have improved the AI score. To rule out writing register, we had Claude Opus 4.8 (the model that wrote the AI reviews) rewrite each of the 134 human reviews into the AI’s register — same substance, AI phrasing and structure (the rewrites kept 97% of the original length and 98% of its citations) — and re-judged with review authorship concealed and using the same model settings as the unmodified-review evaluations. Equalizing style did not help human reviewers: they still lost every decided overall ranking (134 of 134 in both matched and tuned), and, rated independently on the 0–10 individual-rating scale, the restyled human reviews averaged 5.87, close to 5.93 for the unmodified human reviews and far below the AI’s 8.1.

4.8 Undecided Pairs

The AI’s advantage remains large in every review area when the pairs the judge could not decide stay in the denominator as non-wins for the AI. The judge’s rankings were highly consistent across the two presentation orders. Without manuscript access, all 134

matched overall rankings agreed in both presentation orders, so none was discarded; Table 3 reports consistency with and without manuscript access. In the per-area rankings, 6.6% of matched comparisons were order-inconsistent and set aside as undecided; in the tuned (web-enabled) condition this proportion fell to 4.3%. Neither review addressed the area in a further 0.4% of matched comparisons and in none of the comparisons using that per-area selection. These undecided pairs are why the per-area win rates in Figure 1 use decided pairs as their denominators.

When we instead retained all relevant pairs in the denominator, including pairs with order-inconsistent judgments or neither review judged useful, the AI win rates were 89.2% for theory and idea, 99.2% for methods, 100% for weaknesses, 83.8% for framing, 85.4% for communication, 90.9% for figures and tables, and 100% for additional sources (Table 1, panel B). Unresolved pairs contribute to neither side’s wins.

4.9 Accuracy of Citations and Claims

Because reviews can be preferred even if incorrect, we checked accuracy in three ways and found that AI reviews cited no fabricated sources and that 5.5% of their claims about the manuscript were judged incorrect, about a third of the human reviews’ 15.9%.

First, existence. The AI cited 651 sources with a DOI or arXiv identifier. We matched 648 of them automatically to a real record in Crossref or arXiv and confirmed the other three by hand, so all 651 identifiers referred to real published works and none were fabricated (Table 4).

Second, correct use. In a random sample of 20 reviews, containing 112 cited sources in all, a judge from a different model family than the one that wrote the reviews compared each citation against the source’s title and abstract. The cited paper was on-topic for the point the review made in 99 of the 100 cases the judge could decide, with one source off-topic. Most citations simply point the author to relevant work, but in 46 the review attributed a specific finding to a paper, and for these we checked whether the paper’s abstract supported that finding: it held up in 23 of the 24 cases the abstract could settle, leaving a single miscitation in the sample.

Finally, correctness of the reviews’ claims. We checked individual claims, meaning in-

dividual points raised in a review, against the manuscript. This audit covered the matched AI reviews and the human reviews. The AI judge received claims from both reviews in shuffled order under neutral identifiers, with AI/human source labels and previous correctness judgments withheld. A claim counted as incorrect only when the manuscript positively contradicted it; opinions, suggestions, and claims that could not be settled from the manuscript were classified separately as not verifiable. The prompt and manuscript input are given in Supplementary Appendix [D.6](#).

The judge flagged 57/1,039 AI claim assessments (5.5%) and 152/956 human claim assessments (15.9%) as factually incorrect. We also calculated the proportion of claims flagged within each review and averaged those proportions, giving each review equal weight regardless of how many claims it raised. These averages were 5.4% for AI and 15.5% for humans.

At least one claim was flagged in 46/124 AI reviews (37.1%) and 86/134 human reviews (64.2%; Table [5](#)). For AI reviews assessed against two human reviews, these review-level summaries counted the AI review once and treated a claim as flagged if either assessment found it incorrect. AI reviews therefore had fewer flagged errors both among individual claims and across reviews. Nevertheless, just over one-third of AI reviews contained at least one flagged factual error.

A human check of a sample of 90 claims — including every claim the model had flagged as an AI error — agreed with the audit on 87 (96.7%) of them. All three disagreements were cases where the model had wrongly marked a correct AI claim as an error, so the AI’s true error rate is likely below 5.5% and its accuracy advantage over the human reviews holds.

4.10 Detection of Experimentally Induced Factual Errors

To confirm that the AI judge can recognize a statement that contradicts the manuscript, we made two new versions of every review. One version added an accurate sentence about an explicit statement in the manuscript, and the other a minimally altered sentence contradicting it. For each of the 124 manuscripts, we constructed one such pair of sentences about an explicit substantive statement in the paper and used it for that paper’s human

Table 4: **Accuracy of the tuned AI’s citations.** Existence is verified directly against Crossref, doi.org, and arXiv, with no language model involved. On-topic relevance and factual support are judged on a random sample of 20 reviews (112 cited sources). “Cited URLs resolving” excludes addresses that returned bot-block responses, which are inconclusive.

Check	Result
DOI + arXiv identifiers resolving	648 / 651 (99.5%), other 3 confirmed real by hand
Cited URLs resolving	301 / 320 (94%)
Cited sources on-topic (sample)	99% (99 / 100 decided)
Factual attributions supported (sample)	23 / 24

Table 5: **Incorrect claims and reviews containing an error.** A claim is an individual point raised in a review. Incorrect means positively contradicted by the manuscript. This blinded audit covers matched AI and human reviews, with AI/human source labels and previous correctness judgments withheld from the judge. Percentages in parentheses use the displayed denominators.

Review source	Claims flagged incorrect	Reviews with at least one flagged error
AI	57/1,039 (5.5%)	46/124 (37.1%)
Human	152/956 (15.9%)	86/134 (64.2%)

The AI claim denominator counts assessments: ten AI reviews were assessed in two review pairs, giving 1,039 assessments of 959 distinct claims. Reviews are counted once; a claim is flagged if any repeated assessment labels it incorrect. Not-verifiable assessments remain in the claim denominators and are not counted as errors: 316 AI and 523 human.

and matched AI reviews, producing an accurate and a false version of each of 134 human and 124 AI reviews. A separate session checked each contrast against the manuscript and reviews before evaluation. The judge then evaluated all 516 versions independently, given the manuscript but not told which version, or whose review, it was reading, and without the answer key; Supplementary Appendices [D.7–D.9](#) give the prompts for constructing, verifying, and evaluating the versions. The judge flagged all 258 false sentences and none of the 258 accurate ones. On a separate 0–10 factual-reliability scale, versions containing the false statement scored an average of 1.32 points lower for human reviews and 0.80 points lower for AI reviews than their accurate counterparts. This scale measures fidelity to the manuscript, not the review-quality ratings elsewhere in the paper.

4.11 Training-Data Contamination

The AI reviewer’s training data may include work published after a submission, which the human reviewer could not have seen, but the evidence does not support this as a source of AI’s advantage. Submissions span 2019–2026 while the AI reviews were generated in 2026, so if such hindsight helped the AI, its advantage should be largest for the oldest submissions. We find no such gradient (Table 6): splitting the corpus into three submission-year bins, the AI review is preferred in 100% of overall rankings in every bin, wins every review area in every bin, and its 0–10 rating is flat across years (8.2, 8.1, 8.1 for the AI), with no reduction on the most recent (2025–26) submissions, where prior exposure to published work is unlikely.

A direct check of the reviews’ citations points the same way. We scanned every AI review for citations dated after its submission — work the original reviewer could not have consulted. Without web access the AI almost never cites such work: an automated scan flagged 5 of the 124 reviews, only 2 genuine on inspection. With web retrieval enabled it cites recent work far more often, as expected, since that condition can search the web for it. Because the tool-less AI wins just as decisively while citing essentially nothing published after the papers were submitted, its advantage is unlikely to be an artifact of hindsight, although we cannot rule out that knowledge absorbed in training, without being cited, helped it.

4.12 Stochasticity and Model Choice

AI reviews vary stochastically and may depend on the model that writes or judges them, so we replicated the study with a different reviewer prompt and scoring rubric, two AI reviews from each of two model families (Claude Opus and GPT), and judges from both families; the AI review was preferred in 97.1% of 416 comparisons under the Opus judge and 94.2% under the GPT judge. The replication covered 104 human reviews across 99 submissions, and each of its 416 overall rankings set one human review against one of the four AI reviews of the same submission: 404 of the 416 favored the AI under the Opus judge and 392 under the GPT judge. The two reviews from a given model agreed

Table 6: **Robustness to training-data contamination: results by submission-year bin.** All figures are for the *matched* AI condition (the human’s brief, no web retrieval) — the test least able to draw on post-submission hindsight. n : submissions in the bin. *Overall ranking AI win*: share of (AI, human) review pairs in which the blinded judge preferred the AI review overall, with the pair count in parentheses (submissions with two human reviews contribute two pairs, so pairs exceed n). *Per-area ranking AI win*: the same preference, judged within each of the seven reported review areas and pooled over them. *Human rating* and *AI rating*: mean individual rating (0–10) over the seven reported review areas. Human-review ratings are first averaged within each paper, then across papers, so every paper has equal weight. AI contributes one review per paper. The ranking columns retain all review pairs. The AI’s advantage does not reduce for older submissions, contrary to a training-data-hindsight account.

Submission year	n	Overall ranking AI win	Per-area ranking AI win	Human rating	AI rating
≤ 2022	16	100% (18/18)	89%	5.89	8.20
2023–24	53	100% (59/59)	92%	5.94	8.06
2025–26	55	100% (57/57)	89%	5.98	8.10

to within two points on every criterion for every submission. A separate pass that ranked all of a submission’s reviews at once placed the human review last on average under both judges (mean rank 4.7 to 5.0 out of five review types), confirming the result under a second judging method.

4.13 The AI Reviewer’s Prompt and Web Access

The results in Section 3 use the tuned AI reviews; the matched AI reviews perform almost the same, with the exception of winning only 62% of decided pairs in the review area of sources; however, giving AI access to a browser (like the human reviewers had) increased its win rate to 100%, as indicated in Figure 1. Table 7 reports the mean individual ratings for human reviews and both AI conditions in each review area, and Table 8 reports the corresponding AI win rates in per-area rankings, separately for matched and tuned reviews.

4.14 Subgroups of Papers and Authors

AI reviews received higher average ratings than human reviews in every group defined by manuscript novelty, research style (statistical, mathematical, qualitative), faculty rank, career length, and the quality of the journal where the paper was eventually published

Table 7: **Per-review area individual ratings behind Figure 1.** Mean 0–10 score (over the areas judged relevant to each paper) with the number of reviews n it is computed over, for the human review and for the AI review under each condition. Human n can reach 134 while AI n tops out at 124 because 10 of the 124 submissions were independently reviewed by two experts each, whereas the AI wrote a single review per submission. Both drop where an area is not relevant, so additional sources rests on far fewer reviews. Matched is the AI reviewer given no external tools. Tuned is the deployment with web-search enabled, plotted in Figure 1.

Review area	Human (mean, n)	Matched (mean, n)	Tuned (mean, n)
Additional sources	4.73 (99)	5.95 (91)	9.55 (91)
Weaknesses	7.49 (134)	9.05 (124)	9.10 (124)
Methods	7.01 (133)	8.93 (123)	9.02 (123)
Theory & idea	6.97 (130)	8.68 (120)	8.81 (120)
Framing	6.76 (130)	8.63 (120)	8.70 (120)
Communication	5.68 (130)	8.18 (120)	8.18 (120)
Figures & tables	2.95 (132)	7.57 (122)	7.57 (122)
Journal / outlet	2.67 (3)	8.50 (2)	8.50 (2)

Table 8: **Per-area ranking AI win rates behind Figure 1.** For each review area, the AI win rate — the percentage of decided pairs in which the AI review was preferred — with the number of decided pairs n in parentheses, over the areas judged relevant to each paper. Matched is the AI reviewer given no external tools. Tuned is the deployment with web-search enabled, plotted in Figure 1. “Decided” excludes order-inconsistent and unaddressed (“neither”) pairs, so n varies by area and condition. Journal/outlet is relevant for only two submissions and is excluded from Figure 1. Win rates are computed over pairs, of which there are at most 134 (124 submissions plus 10 reviewed by two experts each, while the AI wrote a single review per submission).

Review area	Matched (win %, n)	Tuned (win %, n)
Additional sources	62% (85)	100% (99)
Weaknesses	100% (131)	100% (134)
Methods	98% (130)	100% (132)
Theory & idea	88% (120)	92% (126)
Framing	92% (118)	92% (118)
Communication	92% (120)	92% (120)
Figures & tables	99% (121)	99% (121)
Journal / outlet	100% (3)	100% (3)

(Table 9). We classified submissions by these characteristics using available metadata and compared review ratings within each group. The corpus runs from heavily mathematical formal or game theory, to empirical statistical work, to qualitative and narrative papers; from senior graduate-student authors to full professors, and includes papers that went on to be published in venues up to the top tier of science journals.

Among faculty-rank groups, the largest mean advantage was 2.91 points for pre-faculty or other authors (38 papers), followed by 2.80 for full professors (40 papers). Among publication groups, the largest gap was 2.87 points for other venues or preprints (28 papers), and the smallest was 2.11 for top-general journals (three papers). The AI's rating exceeded the human rating in at least 96% of pairs in every group.

Career length is years since the submitting author's first documented scholarly output, including working papers, conference papers, theses, and datasets. Publication histories were available for 71 papers (74 review pairs). We divided papers into three roughly equal groups based on the author's career length, using a fixed tie-breaking rule to order papers with the same career length. AI reviews scored an average of 2.79 points higher than human reviews in the shortest-career group (24 papers), 2.17 points higher in the middle group (24 papers), and 2.97 points higher in the longest-career group (23 papers). The AI advantage was therefore largest for papers by authors with the longest careers.

5 Journal Policy Choices and Future Research

For two reasons, the policies of scholarly journals (like all policies) are at best informed, but never determined, by scientific findings. First, policies are decided by political processes that are based on factors chosen by the relevant decision makers, in this case the editors, editorial boards, and the commercial or scholarly organizations that own the journals. And second, available empirical evidence, including ours, never exactly aligns with the policy choices to be made. We first state which goal of peer review our evidence addresses, and then describe the dangers of AI reviews followed by five policies journals might adopt; with each policy, we note its risks, what our evidence does not establish, and how future research might better inform the choice.

Table 9: Results by research style, novelty, rank, seniority, and eventual publication. For each subgroup of the corpus, *n pairs* is the number of (AI, human) review pairs; *Human* and *AI* are mean individual ratings (0–10), first averaged over relevant review areas and then over review pairs. For each review area, the AI results use the tuned (web-enabled) condition. *AI win %* is the share of all pairs in which the AI rating exceeds the human rating; ties remain in the denominator. These are rating comparisons, not overall rankings.

Corpus group	<i>n pairs</i>	Human	AI	AI win %
All submissions	134	6.0	8.7	99
Quantitative-statistical	114	6.0	8.6	99
Formal-mathematical	9	6.4	8.8	100
Qualitative	11	5.5	8.9	100
Novelty: lowest tercile	48	5.8	8.7	100
Novelty: middle tercile	43	6.2	8.6	98
Novelty: highest tercile	43	6.1	8.7	100
Submitter: assistant professor	34	6.2	8.6	97
Submitter: associate professor	9	6.4	8.6	100
Submitter: full professor	42	5.9	8.7	100
Submitter: pre-faculty or other	42	5.8	8.7	100
Career length: shortest third	26	5.9	8.6	100
Career length: middle third	24	6.5	8.6	96
Career length: longest third	24	5.6	8.6	100
Published: top general venue	4	6.2	8.3	100
Published: top field venue	9	6.4	8.8	100
Published: solid field venue	13	5.7	8.4	100
Published: other venue or preprint	31	5.9	8.8	100
Not (yet) published	69	6.1	8.7	97

We begin with the stated goals. High quality scholarly reviews are intended to help (1) authors improve their papers and (2) journal editors make publication decisions. Our results speak to (1): For constructive advice to improve scholarly work, our results suggest that AI reviews dominate human reviews at their best, an advantage that seems likely to increase with advances in AI. In addition, since our data are from the high end of the distribution of review quality that journals see in practice, our results probably understate the advantages of AI over human reviewing in practice. It stands to reason that our results would hold for papers by authors outside top research universities, for reviewer pools whose quality varies more than ours, and in disciplines farther outside the social sciences, but establishing these conclusions require new data collection, which we encourage. Our human judges were three former editors who compared ten pairs of reviews without being

asked to read the manuscripts; a larger panel of human editors asked to read the review and manuscript would be a stronger check on the AI judge, if recruitment and compliance proves feasible. Future research would also be useful to show which characteristics of a review lead authors to improve their papers. Our results also do not speak to (2), which is on what basis editors should make publication decisions, but we can use what we learned about (1) to inform journal policy decisions regarding both (1) and (2) as we now describe the dangers and opportunities in new journal policies.

The potential dangers of AI are insidious: Authors brainstorming ideas for a paper with AI have access to more diverse and creative topics than they would without AI, and so are obviously drawn to use it, but their competitors brainstorming in the same area, even using different AI models, wind up with roughly the same ideas, and all will tend toward the conventional, orthodox, and majoritarian ideas (Hao et al., 2026; Wenger and Kenett, 2026). In other words, AI benefits us individually but risks making the collective efforts of science more homogenous and ordinary. Algorithms exist to ameliorate this problem, but they would need to be used (e.g., Luo et al., 2026). AI reviews may also have some of the same issues, potentially pushing individual papers and science as a whole toward homogenous arguments and conclusions. The heterogeneity of human reviewers, who disagree with each other so often, can thus be both advantageous and damaging, whereas the small number of AIs used for reviewing may take the whole literature closer to a single point of failure, controlled by a few technology companies. The potential of AI biasing journal acceptance decisions toward certain hypotheses, questions, or styles is real, but the process would be easier to study and adjust than with human reviewers.

Next, if these dangers of AI reviewing can be constrained, then the opportunities could be substantial; we describe these in the context of five possible journal policy changes. First, some journals may begin to recommend or require that authors iterate with AI before submission to improve their papers (via what are essentially self-supervised revise-and-resubmits), and continuing until authors are convinced they have resolved all resolvable issues and improved the paper as much as possible within time or resource constraints. Our evidence on this policy is indirect: The judges in our study were an AI model and

former journal editors, not the authors who would act on the advice, and the authors rated only the human review they received, not the AI review. Asking authors to compare the two reviews of their own paper would be valuable research, probably easiest to arrange as a routine part of a journal’s workflow.

Second, journals may consider adopting and publishing a specific AI prompt for authors to use for evaluation of their papers (for a starting point, editors could combine the *rapidPeer* prompt in Supplementary Appendix B with the current “instructions to contributors” each journal already makes public). This policy could then use AI to better align submissions with each journal’s specific editorial standards, reduce uncertainty in the review process, improve quality, and potentially even reduce rejections, especially desk rejects. Given the speed of progress in AI, continuing research would be valuable to see how our results change with new foundation models, better prompts, open source and commercial review-specific AI services, AI access to a paper’s replication dataset, and novel combinations of AI and human reviews (Meng, 2026). Research is also needed on whether our results hold with the less capable models and default settings many will use.

Third, journals could require an AI review, generated by the journal’s official AI prompt, be submitted with the paper, as well as a letter from the author explaining any remaining open questions or disagreements with AI. The author’s letter, which is akin to the “response to reviewers” authors are already often asked to write, is crucial so that AI model creators do not get to determine the content of scholarly papers and because, as the paper improves during the iterations, AI sometimes becomes absurdly pedantic even if still technically correct (such as “yes $2 + 2 = 4$, but you never defined the $+$ sign”). Our judges, human and AI, did not know which (if any) review was written by AI, but under this policy editors and authors would, and so future research to repeat our comparison with disclosed sources may be useful.

Fourth, some journals may wish to go farther by ending human reviews entirely and using AI reviews instead, as our evidence shows that AI reviews give more constructive advice, and AI reviews alone have been shown to improve quality (Thakkar et al., 2026). However, if the journal’s only source of information is an AI review that authors can run

themselves in advance, authors will revise until that review is favorable, and the review will then tell the editor little the author has not already addressed. Indeed, this decision may already be a fait accompli: despite the efforts of some editors to stop AI use in (supposedly human) reviews, AI is increasingly used by human reviewers (Naddaf, 2026), and we informally have observed use by editors increasing too. At the 2026 International Conference on Machine Learning, which randomly assigned reviewers to policies on AI use, 22.5% of the surveyed reviewers assigned to a policy forbidding AI reported using it anyway (Kim et al., 2026). We note that our results are estimates of quantities at the level of the review, and so we would all benefit from future research on the collective consequences of AI review policies, such as the likely equilibrium resulting from the combination of AI reviewing, AI writing, AI journal decision-making, or combined AI-assisted versions of each.

Finally, a more realistic policy change that some journals may adopt is to skip the *routine* use of human reviews. Then the editor could review a dossier submitted by the author (consisting of the paper, the AI review generated from the journal’s official AI prompt, and the author’s response to the AI), a new AI review generated by prompts the editor chooses and does not share publicly (such as including manipulation checks; Murray 2025), and, if necessary for particular cases, a review by a chosen editorial board member or even, in difficult cases, an outside traditional human reviewer. The editor would then make a final publication decision. We compared whole reviews, and a human review that loses overall may still raise a useful point the AI review misses.

Some version of these ideas has the potential to greatly speed the publication process, enabling decisions within days rather than months. If the dangers can be avoided, it may remove the bias and painfully adversarial parts of the process for the author (such as having to make unnecessary or sycophantic changes in papers and revise-and-resubmit responses to avoid offending a human reviewer). Papers submitted using iterative AI review will likely be more free of errors and weaknesses in logic, math, references, argument, scholarly literature, data analyses, graphic presentations, and even journal stylistic choices (AI makes mistakes too, although, in our study, only about a third as often as

human reviewers). Problems of special concern to each journal could be more easily avoided by including them in the journal's official AI prompt (as AI is considerably better at following long lists of detailed "instructions to authors" than humans).

If we can get the policies right, the job of the editor may then be improved and of more interest to the high quality people the scientific community attracts for this job. Some use of AI may also help with the framing, impact, overall importance of the paper, and suitability for the journal, but these crucial decisions are the most important ones and so journals will presumably wish these to be left to the journal editor. Even if all AI dangers are avoided, authors are conscientious in iterating with AI to improve their paper, and paper submissions are all correct or close to it, whether a correct paper meets the standards of any one scholarly journal should presumably only be decided by a human journal editor with the proper authority and responsibility.

Any policy ideas such as these carry additional risks that journals will need to manage. Authors who revise until the published prompt's AI review is favorable, as raised above, may also change their papers in ways the AI rewards that have nothing to do with the quality of the science. Some partial protections are already included — the author's letter answering the AI review, a second AI review from a prompt the editor does not disclose, and manipulation checks — but how they work in practice is an empirical question that needs future research. Thus, we propose that an important component of adopting AI at any journal should be an ongoing study of authors' disagreements with AI (with the journal editor's views) so we can continuously improve by updating their disclosed and undisclosed prompts and choosing among AI models.

To conclude, we ask: is our title overstated? One AI says yes, others say no. We're sticking with it!

Ethics Statement

This research was approved by our university IRB. *RapidPeer* terms of use promise confidentiality but allow papers and reviews to be used for further research. All use of AI in this paper was under enterprise licensing agreements that prevent companies from training

on user data. Because we conceived of and run *RapidPeer*, we have a conflict of interest, but our empirical results go against these interests.

References

- Aczel, Balazs, Ann-Sophie Barwich, Amanda B. Diekman, Ayelet Fishbach, Robert L. Goldstone, Pablo Gomez, Odd Erik Gundersen, Paul T. von Hippel, Alex O. Holcombe, Stephan Lewandowsky, Nazbanou Nozari, Franco Pestilli, and John P. A. Ioannidis (2025). “The present and future of peer review: Ideas, interventions, and evidence”. In: *Proceedings of the National Academy of Sciences* 122.5. DOI: [10 . 1073/pnas.2401232121](https://doi.org/10.1073/pnas.2401232121).
- Aczel, Balazs, Barnabas Szaszi, and Alex O. Holcombe (2021). “A billion-dollar donation: estimating the cost of researchers’ time spent on peer review”. In: *Research Integrity and Peer Review* 6.1, p. 14. DOI: [10.1186/s41073-021-00118-2](https://doi.org/10.1186/s41073-021-00118-2).
- Anthropic (2026). *Claude Opus 4.8*. <https://www.anthropic.com/claude/opus>. Large language model; model documentation, accessed 2026.
- Armstrong, J. Scott (1997). “Peer review for journals: Evidence on quality control, fairness, and innovation”. In: *Science and Engineering Ethics* 3.1, pp. 63–84. DOI: [10 . 1007/s11948-997-0017-3](https://doi.org/10.1007/s11948-997-0017-3).
- Biswas, Joydeep, Sheila Schoepp, Gautham Vasan, Anthony Opiari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins (2026). “AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot”. In: *arXiv preprint arXiv:2604.13940*. DOI: [10.48550/arXiv.2604.13940](https://doi.org/10.48550/arXiv.2604.13940).
- Bornmann, Lutz, Rüdiger Mutz, and Hans-Dieter Daniel (2010). “A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants”. In: *PLoS ONE* 5.12, e14331. DOI: [10.1371/journal.pone.0014331](https://doi.org/10.1371/journal.pone.0014331).
- Chong, Sin Wang and Tingjun Lin (2024). “Feedback practices in journal peer-review: a systematic literature review”. In: *Assessment & Evaluation in Higher Education* 49.1, pp. 1–12. DOI: [10.1080/02602938.2022.2164757](https://doi.org/10.1080/02602938.2022.2164757).
- Doskaliuk, Bohdana, Olena Zimba, Marlen Yessirkepov, Iryna Klishch, and Roman Yatsyshyn (2025). “Artificial Intelligence in Peer Review: Enhancing Efficiency While Preserving Integrity”. In: *Journal of Korean Medical Science* 40.7, e92. DOI: [10 . 3346/jkms.2025.40.e92](https://doi.org/10.3346/jkms.2025.40.e92).
- Farber, Shai (2025). “Comparing human and AI expertise in the academic peer review process: towards a hybrid approach”. In: *Higher Education Research & Development* 44.4, pp. 871–885. DOI: [10.1080/07294360.2024.2445575](https://doi.org/10.1080/07294360.2024.2445575).
- Fichtl, Alexander M., Lukas Ellinger, Josefin Kelber, Kryštof Olík, and Georg Groh (2026). “AI-Assisted Peer Review Across Research Communities: From Reviewer AI Policies to LLM Review Quality”. In: *arXiv preprint arXiv:2608.03581*. DOI: [10 . 48550/arXiv.2608.03581](https://doi.org/10.48550/arXiv.2608.03581).

- Gasparyan, Armen Yuri and George D. Kitas (2012). “Best peer reviewers and the quality of peer review in biomedical journals”. In: *Croatian Medical Journal* 53.4, pp. 386–389. DOI: [10.3325/cmj.2012.53.386](https://doi.org/10.3325/cmj.2012.53.386).
- Hao, Qian Yue, Fengli Xu, Yong Li, and James Evans (2026). “Artificial intelligence tools expand scientists’ impact but contract science’s focus”. In: *Nature* 649.8099, pp. 1237–1243. DOI: [10.1038/s41586-025-09922-y](https://doi.org/10.1038/s41586-025-09922-y).
- Hosseini, Mohammad and Serge P. J. M. Horbach (2023). “Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review”. In: *Research Integrity and Peer Review* 8.1, p. 4. DOI: [10.1186/s41073-023-00133-5](https://doi.org/10.1186/s41073-023-00133-5).
- Kim, Sunnie S. Y., Wesley Hanwen Deng, Jennifer Wortman Vaughan, Buxin Su, Weijie Su, Alekh Agarwal, Sharon Li, Martin Jaggi, Daniel G. Goldstein, Nihar B. Shah, and Miroslav Dudík (2026). “Use and Effects of LLMs in Peer Review: A Randomized Experiment and Survey at ICML 2026”. In: *arXiv preprint arXiv:2609.19420*. DOI: [10.48550/arXiv.2609.19420](https://doi.org/10.48550/arXiv.2609.19420).
- Laurito, Walter, Benjamin Davis, Peli Grietzer, Tomáš Gavenčíak, Ada Böhm, and Jan Kulveit (2025). “AI–AI bias: Large language models favor communications generated by large language models”. In: *Proceedings of the National Academy of Sciences* 122.31, e2415697122. DOI: [10.1073/pnas.2415697122](https://doi.org/10.1073/pnas.2415697122).
- Liang, Weixin, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou (2024). “Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis”. In: *NEJM AI* 1.8. DOI: [10.1056/AIoa2400196](https://doi.org/10.1056/AIoa2400196). URL: <https://ai.nejm.org/doi/full/10.1056/AIoa2400196>.
- Lin, Zhicheng (2025). “Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review”. In: *arXiv preprint arXiv:2507.06185*. DOI: [10.48550/arXiv.2507.06185](https://doi.org/10.48550/arXiv.2507.06185).
- Luo, Queenie, Gary King, Michael Puett, and Michael D. Smith (2026). “Inducing Sustained Creativity and Diversity in Large Language Models”. In: URL: <https://GaryKing.org/quest>.
- Macharla Vasu, Sai Suresh, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, and Mario Fritz (2026). “Justice in Judgment: Unveiling (Hidden) Bias in LLM-assisted Peer Reviews”. In: *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 307–330. URL: <https://aclanthology.org/2026.findings-acl.14/>.
- Meng, Xiao-Li (2026). “Comments: Can Statistics and AI Technologies Help Our Troubled Review System?” In: *Journal of the American Statistical Association* 121.554, pp. 855–861. DOI: [10.1080/01621459.2026.2657554](https://doi.org/10.1080/01621459.2026.2657554).
- Murray, Toby (2025). “PhantomLint: Principled Detection of Hidden LLM Prompts in Structured Documents”. In: *arXiv preprint arXiv:2508.17884*. DOI: [10.48550/arXiv.2508.17884](https://doi.org/10.48550/arXiv.2508.17884).
- Mutz, Rüdiger, Lutz Bornmann, and Hans-Dieter Daniel (2012). “Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and Its Determinants: A General Estimating Equations Approach”. In: *PLoS ONE* 7.10, e48509. DOI: [10.1371/journal.pone.0048509](https://doi.org/10.1371/journal.pone.0048509).

- Naddaf, Miryam (2025). “AI is transforming peer review — and many scientists are worried”. In: *Nature* 639, pp. 852–854. DOI: [10.1038/d41586-025-00894-7](https://doi.org/10.1038/d41586-025-00894-7).
- (2026). “More than half of researchers now use AI for peer review — often against guidance”. In: *Nature* 649, pp. 273–274. DOI: [10.1038/d41586-025-04066-5](https://doi.org/10.1038/d41586-025-04066-5).
- OpenAI (2026). *GPT-5.5*. <https://openai.com/index/gpt-5/>. Large language model; model documentation, accessed 2026.
- Porsdam Mann, Sebastian, Mateo Aboy, Joel Jiehao Seah, Zhicheng Lin, Xufei Luo, Daniel Rodger, Hazem Zohny, Timo Minssen, Julian Savulescu, and Brian D. Earp (2025). “AI and the Future of Academic Peer Review”. In: *arXiv preprint arXiv:2509.14189*. DOI: [10.48550/arXiv.2509.14189](https://doi.org/10.48550/arXiv.2509.14189).
- Rajakumar, Hamrish Kumar, Kailash Abhishek Sankaran, Manasi Pillai Ashok, and Srinivas Rachoori (2026). “Peer review in the age of artificial intelligence: a comparative study of human and AI-generated review reports”. In: *Postgraduate Medical Journal* 102.1208, pp. 551–559. DOI: [10.1093/postmj/qgag005](https://doi.org/10.1093/postmj/qgag005).
- Russo Latona, Giuseppe, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West (2024). “The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates”. In: *arXiv preprint arXiv:2405.02150*. Published in Proceedings of the ACM on Human–Computer Interaction. DOI: [10.48550/arXiv.2405.02150](https://doi.org/10.48550/arXiv.2405.02150).
- Shin, Hyungyu, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim (2025). “Mind the Blind Spots: A Focus-Level Evaluation Framework for LLM Reviews”. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35630–35656. URL: <https://aclanthology.org/2025.emnlp-main.1805/>.
- Tennant, Jonathan P. and Tony Ross-Hellauer (2020). “The limitations to our understanding of peer review”. In: *Research Integrity and Peer Review* 5.1, p. 6. DOI: [10.1186/s41073-020-00092-1](https://doi.org/10.1186/s41073-020-00092-1).
- Thakkar, Nitya, Mert Yuksekogul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou (2026). “A large-scale randomized study of large language model feedback in peer review”. In: *Nature Machine Intelligence* 8, pp. 326–336. DOI: [10.1038/s42256-026-01188-x](https://doi.org/10.1038/s42256-026-01188-x).
- Wenger, Emily and Yoed N Kenett (2026). “Large language models are homogeneously creative”. In: *PNAS Nexus* 5.3. <https://doi.org/10.1093/pnasnexus/pgag042>, pgag042. DOI: [10.1093/pnasnexus/pgag042](https://doi.org/10.1093/pnasnexus/pgag042).
- Wright, Duncan E. (2024). “Five problems plaguing publishing in the life sciences—and one common cause”. In: *FEBS Letters* 598.18, pp. 2227–2239. DOI: [10.1002/1873-3468.15018](https://doi.org/10.1002/1873-3468.15018).
- Ye, Rui, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen (2024). “Are We There Yet? Revealing the Risks of Utilizing Large Language Models in Scholarly Peer Review”. In: *arXiv preprint arXiv:2412.01708*. DOI: [10.48550/arXiv.2412.01708](https://doi.org/10.48550/arXiv.2412.01708).
- Zancanaro, Edoardo, Andreas Giannopoulos, Alessia Gimelli, and Karl-Patrik Kresoja (2026). “AI as a peer reviewer: a blinded comparative study of LLM-generated and human reviews in a cardiology journal”. In: *European Heart Journal – Imaging Methods and Practice* 4.1, qyag097. DOI: [10.1093/ehjimp/qyag097](https://doi.org/10.1093/ehjimp/qyag097).