

On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists

Seungone Kim

seungonk@andrew.cmu.edu

Carnegie Mellon University <https://orcid.org/0009-0003-0479-9550>

Dongkeun Yoon

KAIST

Kiril Gashteovski

NEC Laboratories Europe & Ss. Cyril and Methodius University in Skopje

Juyoung Suk

KAIST

Jinheon Baek

KAIST

Pranjal Aggarwal

Carnegie Mellon University

Ian Wu

Carnegie Mellon University

Viktor Zaverkin

INM - Leibniz Institute for New Materials; Saarland University, Department of Computer Science;
German Research Center for Artificial Intelligence (DFKI)

Spase Petkoski

<https://orcid.org/0000-0003-4540-6293>

Daniel Schrider

Department of Genetics, University of North Carolina at Chapel Hill

Ilija Dukovski

Boston University & Ss. Cyril and Methodius University in Skopje

Francesco Santini

Basel Muscle MRI, Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland;
Department of Radiology, University Hospital of Basel, Basel, Switzerland

Biljana Mitreska

Department of Physics and Astronomy, University of Manchester, Manchester, United Kingdom

Yong Jeong

Korea Advanced Institute of Science and Technology <https://orcid.org/0000-0002-5907-3787>

Kyeongha Kwon

Korea Advanced Institute of Science & Technology (KAIST) <https://orcid.org/0000-0002-2373-7799>

Youngmin Sim

KAIST

Dragana Manasova

Massachusetts Institute of Technology

Arthur Porto

Florida Museum of Natural History, University of Florida

Biljana Mojsoska

Department of Science and Environment, Roskilde University

Makoto Takamoto

NEC Laboratories Europe

Marko Shuntov

Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen, Denmark

Ruoqi Liu

Stanford University

Hyunjoo Lee

Seoul National University <https://orcid.org/0000-0001-9963-2515>

Niyazi Dinç

École Polytechnique Fédérale de Lausanne

Yehhyun Jo

Institute for Basic Science

Sunkyu Han

Korea Advanced Institute of Science and Technology <https://orcid.org/0000-0002-9264-6794>

Chungwoo Lee

KAIST

Huishan Li

KAIST

Esther Tsai

Brookhaven National Laboratory <https://orcid.org/0000-0002-5885-6317>

Ergun Simsek

University of Maryland Baltimore County <https://orcid.org/0000-0001-9075-7071>

Khushboo Shafi

KAIST

Yeonseung Chung

Korea Advanced Institute of Science and Technology <https://orcid.org/0000-0002-1501-2747>

Jihye Park

Lawrence Berkeley National Laboratory

Aleksandar Shulevski

The Netherlands Institute for Radio Astronomy & Ss. Cyril and Methodius University in Skopje
<https://orcid.org/0000-0002-1827-0469>

Henrik Christiansen

NEC Laboratories Europe

Yoosang Son

KAIST

Elly Knight

Department of Biological Sciences, University of Alberta

Amanda Montoya

The University of Texas MD Anderson Cancer Center

Jeongyoun Ahn

KAIST

Christian Langkammer

Medical University of Graz

Heera Moon

KAIST

Changwon Yoon

KAIST

Nikola Stikov

Polytechnique Montréal & Montreal Heart Institute & Ss. Cyril and Methodius University in Skopje

Mooseok Jang

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology
<https://orcid.org/0000-0003-1977-9539>

Edward Choi

KAIST

Junhan Kim

KAIST <https://orcid.org/0000-0002-4274-9373>

Yeon Sik Jung

Korea Advanced Institute of Science and Technology (KAIST) <https://orcid.org/0000-0002-7709-8347>

Woo Youn Kim

KAIST <https://orcid.org/0000-0001-7152-2111>

Jae Kyoung Kim

KAIST <https://orcid.org/0000-0001-7842-2172>

Ishraq Anjum

University of Maryland Baltimore County

Hyun Uk Kim

Korea Advanced Institute of Science and Technology (KAIST) <https://orcid.org/0000-0001-7224-642X>

Drew Bridges

Carnegie Mellon University

Carolyn Lawrence

NEC Laboratories Europe

Xiang Yue

CMU <https://orcid.org/0000-0003-4547-1685>

Alice Oh

KAIST

Akari Asai

University of Washington <https://orcid.org/0000-0002-1055-6657>

Sean Welleck

Carnegie Mellon University

Graham Neubig

CMU

Social Sciences - Article

Keywords:

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10896922/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

There is **NO** Competing Interest.

Figure 9 is not available with this version

On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists

Seungone Kim¹ Dongkeun Yoon² Kiril Gashteovski^{3,4} Juyoung Suk² Jinheon Baek²
Pranjal Aggarwal¹ Ian Wu¹ Viktor Zaverkin⁵ Spase Petkoski^{4,6} Daniel R. Schrider⁷
Ilija Dukovski^{4,8} Francesco Santini^{9,10} Biljana Mitreska¹¹ Yong Jeong² Kyeongha Kwon²
Youngmin Sim² Dragana Manasova¹² Arthur Porto¹³ Biljana Mojsoska¹⁴ Makoto Takamoto³
Marko Shuntov¹⁵ Ruoqi Liu¹⁶ Hyunjoo Jenny Lee² Niyazi Ulas Dinç¹⁷ Yehhyun Jo¹⁸
Sunkyu Han² Chungwoo Lee² Huishan Li² Esther H. R. Tsai¹⁹ Ergun Simsek²⁰
Khushboo Shafi² Yeonseung Chung² Jihye Park²¹ Aleksandar Shulevski^{4,22}
Henrik Christiansen³ Yoosang Son² Elly Knight²³ Amanda Montoya²⁴
Jeongyoun Ahn² Christian Langkammer²⁵ Heera Moon² Changwon Yoon²
Nikola Stikov^{4,26,27} Mooseok Jang² Edward Choi² Junhan Kim² Yeon Sik Jung²
Woo Youn Kim² Jaekyoung Kim^{2,18} Ishraq Md Anjum²⁰ Hyun Uk Kim² Drew Bridges¹
Carolyn Lawrence³ Xiang Yue¹ Alice Oh² Akari Asai¹ Sean Welleck¹ Graham Neubig¹

¹Carnegie Mellon University ²KAIST ³NEC Laboratories Europe

⁴Ss. Cyril and Methodius University in Skopje ⁵INM - Leibniz Institute for New Materials;

Saarland University German Research Center for Artificial Intelligence (DFKI)

⁶Aix Marseille University, INSERM ⁷University of North Carolina at Chapel Hill ⁸Boston University

⁹University of Basel ¹⁰University Hospital of Basel ¹¹University of Manchester

¹²Massachusetts Institute of Technology ¹³Florida Museum of Natural History, University of Florida

¹⁴Roskilde University ¹⁵University of Copenhagen ¹⁶Stanford University

¹⁷École Polytechnique Fédérale de Lausanne ¹⁸Institute for Basic Science (IBS)

¹⁹Brookhaven National Laboratory ²⁰University of Maryland Baltimore County

²¹Lawrence Berkeley National Laboratory ²²The Netherlands Institute for Radio Astronomy

²³University of Alberta ²⁴The University of Texas MD Anderson Cancer Center

²⁵Medical University of Graz ²⁶Polytechnique Montréal ²⁷Montreal Heart Institute

Abstract

As AI capabilities advance, AI reviewers (agents that read a manuscript and produce a referee report) are beginning to be deployed in scientific peer review. Yet, their credibility remains in question: many scientists assume that they cannot judge scientific validity, while others are more optimistic about their readiness. In this work, we conduct a study in which 45 scientists in Physical, Biological, and Health Sciences spend 469 hours rating 2,960 individual criticisms (each targeting one aspect of a paper) from human-written and AI-generated reviews of 82 *Nature*-family papers, scoring each on correctness, significance, and evidence sufficiency. On a composite rating of reviews across all three dimensions, an AI reviewer powered by GPT-5.2 scores above each paper’s top-rated human reviewer (60.0% vs. 48.2% of criticisms, $p = 0.009$). Also, AI reviewers’ correct criticisms are more often rated significant, and surface a distinct 26% of legitimate issues that no human raises. However, AI reviewers are more often incorrect, overlap with each other far more than humans do with each other, and exhibit 16 weaknesses rare in human reviews. Overall, our results position current AI reviewers as complements to, not substitutes for, human reviewers.

1 Introduction

Peer review has long served as the cornerstone of the scholarly publication system, ensuring the credibility, rigor, and cumulative advancement of scientific knowledge [Gannon, 2001, Kelly et al., 2014, Siler et al., 2015]. Effective review requires criticisms that are factually correct, that identify problems consequential enough to matter, and that are supported by evidence an author can act on. That standard is becoming harder to meet, as submissions have grown sharply in recent years while the pool of reviewers with the necessary time and subfield expertise has not, and researchers increasingly report declining review quality [Chen et al., 2025, Powell, 2016]. AI agents that generate paper reviews, which we refer to as *AI reviewers* [Liu and Shah, 2023, Kuznetsov et al., 2024], are one response now deployed at scale: the AI conference AAAI-26 ran them on all 22,977 main-track submissions [Biswas et al., 2026], and the medical journal *NEJM AI* has built them into its “Fast Track” process [Manrai et al., 2025].

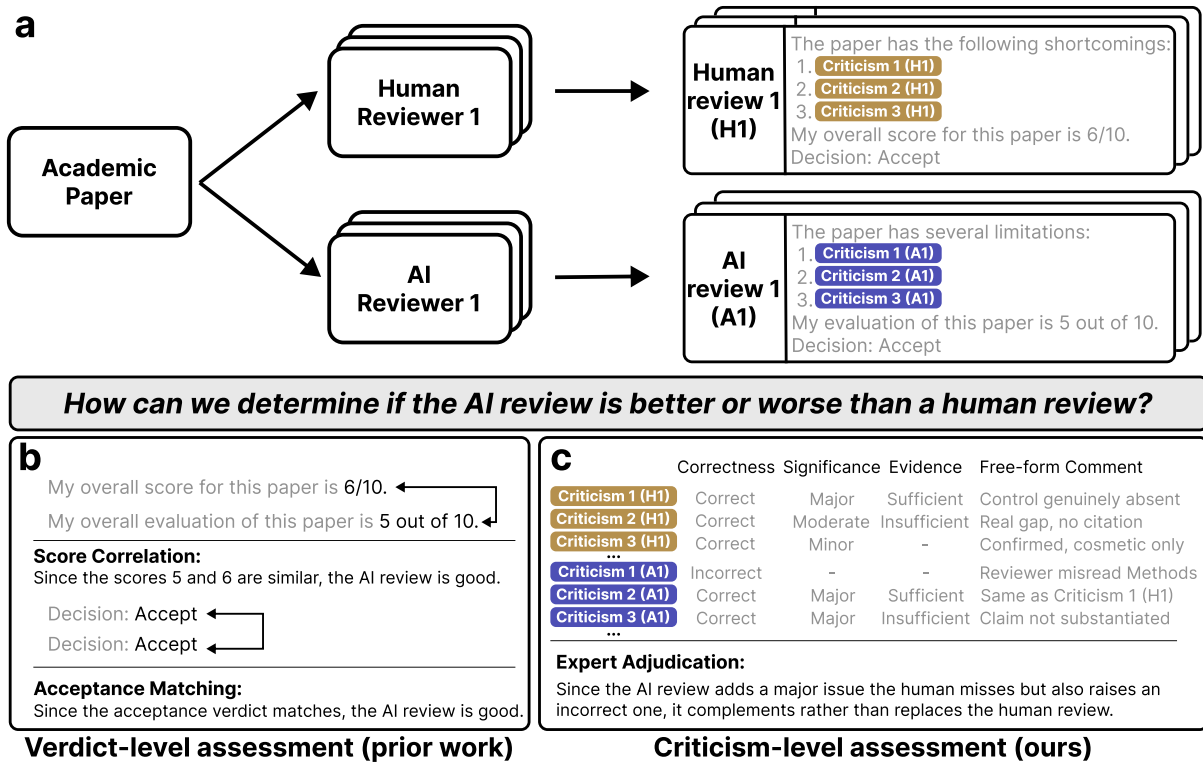


Figure 1: **Motivation of our evaluation protocol.** **a**, Human and AI reviewers produce reviews of the same paper. **b**, Prior work assesses AI reviews at the review level, using score correlation and acceptance matching. Similar scores and matching recommendations do not establish that an AI review contains useful feedback, and neither indicates whether the underlying criticisms agree. **c**, We assess reviews at the criticism level, taking the individual criticism as the unit of analysis. A domain expert adjudicates every criticism from both sources on correctness, significance, and evidence sufficiency, and records a free-form comment justifying the adjudication. This protocol supports the three analyses that follow: per-criterion and composite performance against human reviewers (§ 2), overlap between AI and human reviews (§ 3), and a categorization of AI reviewer strengths and weaknesses drawn from the expert comments (§ 4). Criticisms are labelled by source, H for human reviewers and A for AI reviewers.

However, these deployments are running ahead of the evidence. To date, prior work has assessed the quality of AI reviews by measuring whether AI reviewers reach the same verdict as humans, comparing overall scores, recommendations, and holistic ratings (Figure 1b) [Saad et al., 2024, Zhu et al., 2025, Idahl and Ahmadi, 2025, Zhang et al., 2026, Lu et al., 2026]. This target is weaker than it appears: the consistency experiments that were conducted by the *NeurIPS* conference (an AI conference) found that roughly half of the papers accepted by one committee were rejected by an independent second committee [Cortes and Lawrence, 2021, Beygelzimer et al., 2023], which makes the reference itself substantially arbitrary. More fundamentally, two reviews can reach an identical recommendation while sharing no common criticisms. Hence, verdict-level assessment does not speak to how the criticisms in an AI review compare in quality with those of the human referees on the same paper, whether the two groups identify the same problems, and what AI reviewers concretely do well and badly.

To address this, we introduce a criticism-level assessment framework that compares reviews by the criticisms they contain rather than the verdicts they reach (Figure 1c). The framework decomposes each review into individual *criticisms*, atomic statements that each target one aspect of a paper, which makes it possible to ask whether two reviews raise the same issues and not merely how good each is. Since feedback quality depends on conventions specific to the paper’s subfield, a domain scientist expert in the paper’s subfield then adjudicates every criticism on three dimensions (correctness, significance, evidence sufficiency), together with a free-form comment justifying the judgment. Each annotator judges six reviews per paper, three human referee reports from the published transparent peer review and three AI reviews generated by our pipeline. Also, they rank the human reviews against one another, giving two reference points per paper so that AI reviews are compared against both the strongest and the weakest human review the paper received. Applying this framework, we pair 82 *Nature*-family papers with their referee reports and with reviews from GPT-5.2, Claude Opus 4.5, and Gemini 3.0 Pro, and forty-five domain scientists spanning Physical, Biological, and Health Sciences collectively spend 469 hours adjudicating 2,960 criticisms.

Reviewer	Correctness (%) mean [95% CI]	Significance (0–2) mean [95% CI]	Evidence (%) mean [95% CI]	n_{items}	n_{papers}
Top-Rated Human	92.3 [89.2, 95.0]	1.39 [1.30, 1.49]	92.2 [88.5, 95.3]	1,139	82
Lowest-Rated Human	79.1 [73.0, 84.6]	1.30 [1.18, 1.42]	89.7 [84.7, 94.0]	833	82
GPT-5.2	86.2 [80.7, 91.2]	1.61 [1.50, 1.70]	97.1 [93.7, 99.5]	442	81
vs Top-Rated Human	$d = -0.225$, $p = .046^*$	$r = +0.491$, $p < .001^*$	$d = +0.233$, $p = .041^*$		
vs Lowest-Rated Human	$d = +0.220$, $p = .051$	$r = +0.562$, $p < .001^*$	$d = +0.234$, $p = .048^*$		
Claude Opus 4.5	83.7 [78.2, 88.6]	1.53 [1.43, 1.63]	96.5 [93.1, 99.1]	475	81
vs Top-Rated Human	$d = -0.337$, $p = .003^*$	$r = +0.301$, $p = .028^*$	$d = +0.243$, $p = .035^*$		
vs Lowest-Rated Human	$d = +0.134$, $p = .230$	$r = +0.434$, $p = .003^*$	$d = +0.274$, $p = .022^*$		
Gemini 3.0 Pro	81.9 [76.5, 87.1]	1.56 [1.46, 1.65]	89.5 [84.0, 94.2]	460	82
vs Top-Rated Human	$d = -0.424$, $p < .001^*$	$r = +0.423$, $p = .003^*$	$d = -0.099$, $p = .380$		
vs Lowest-Rated Human	$d = +0.088$, $p = .428$	$r = +0.460$, $p = .001^*$	$d = -0.012$, $p = .918$		

Table 1: **AI reviewers identify more significant issues but with lower factual correctness than the top-rated human reviewers.** Paper-level means with 95% bootstrap CIs. Shaded rows: pairwise effect sizes and p -values versus each human reference (positive = AI higher; d = Cohen’s d for binary metrics, r = rank-biserial for ordinal significance). $^*p < 0.05$. See Appendix B for item-level rates and a GLMM robustness analysis.

On the composite of all three dimensions, GPT-5.2 exceeds each paper’s top-rated human referee (60.0% vs. 48.2%, $p = 0.009$), and all three AI reviewers outperform the lowest-rated human reviewer (all $p < 0.01$). Thus, according to our criteria, current AI reviewers are already superior to the lowest quality review a paper is likely to receive. Furthermore, the composite conceals a tradeoff that the per-dimension results make visible: AI reviewers are less often correct than the top-rated human (81.9 to 86.2% against 92.3%), but their correct criticisms are more often judged to be significant (§ 2).

Individual review quality, however, is not the same as the overall contribution a review makes as part of the review process as a whole. To quantify this, we measure whether each criticism raised by one reviewer is also raised by another. A single AI reviewer picks up 27.1% of the criticisms a human referee raises, barely above the 25.8% a second human picks up, so an AI review resembles a human review about as much as two human reviews resemble each other. What falls outside that overlap is largely sound: 26.0% of AI criticisms have no counterpart in any human review, and 81.8% of those are correct, meaning that AI reviewers contribute observations that human reviewers do not make, rather than simply producing noise. Replacing an entire panel is a different matter: two human referees share only 3.4% of their criticisms, whereas two AI reviewers share 20.9%. Therefore, substituting one AI reviewer for one human therefore leaves a panel’s diversity nearly intact, but a panel composed entirely of AI reviewers lacks diversity (§ 3).

The free-form comments explain what drives these differences. Domain experts flagged 260 weaknesses and 132 strengths in AI reviews, and the weaknesses are typically contextual rather than factual: AI reviewers misjudge severity of flaws against field-specific norms, and assert that something is missing when the paper in fact supplies it in another section or the supplementary material. Their strengths are the converse, namely forms of scrutiny that reward diligence over judgment, such as checking statistical assumptions and reading the submitted source code (§ 4). Together these results position current AI reviewers as complements to, not substitutes for, human reviewers, and locate the gap in subfield knowledge rather than in reasoning or thoroughness. We release the annotation corpus of 2,960 expert adjudications with written justifications, and deploy the review pipeline as a public service at cmu-paper-reviewer.com.

2 In which aspects are AI reviewers better or worse than human reviewers?

Expert annotation setup We study 82 *Nature*-family papers whose referee reports are published under transparent peer review, spanning the physical, biological, and health sciences (paper selection criteria are in § 6). Each paper is assigned to at least one domain scientist working in its subfield, with 45 scientists producing 109 assignments; 27 papers are annotated independently by two scientists, which lets us measure inter-annotator agreement. Each annotator provides three kinds of annotation over the paper’s six reviews, three written by human referees and three generated by AI reviewers. First, they rate every *review item*, an atomic criticism targeting one aspect of a paper, from all six reviews. The three dimensions are applied in sequence: whether the item is factually *correct* about the paper (binary); if correct, how *significant* the issue it raises is (not significant, marginally significant, significant); and if at least marginally significant, whether the *evidence* supporting it is sufficient for an author to act on (binary). Each rating carries a free-form comment justifying the judgment. In total, this yields 2,960 annotated review items

Reviewer	Per-group “fully positive” rate		Fully Positive	Total Items	Papers
	Paper-level mean [95% CI]	Item-level rate [95% CI]			
Top-Rated Human	48.2 [42.2, 54.3]	47.8 [44.9, 50.7]	544	1,139	82
Lowest-Rated Human	36.2 [30.0, 42.4]	29.3 [26.3, 32.5]	244	833	82
GPT-5.2	60.0 [52.3, 67.4]	58.6 [54.0, 63.1]	259	442	81
Claude Opus 4.5	53.1 [45.6, 60.7]	54.7 [50.2, 59.2]	260	475	81
Gemini 3.0 Pro	50.2 [42.7, 57.7]	51.3 [46.7, 55.8]	236	460	82

Pairwise paired differences (paper-level, row – column; paired *t*-test)

Comparison	Diff	95% CI (diff)	Cohen’s <i>d</i>	<i>p</i> -value	<i>n</i> _{paired}
Top-Rated H. vs Lowest-Rated H.	+12.1%	[+5.3, +18.9]	+0.39	0.0007***	82
Top-Rated H. vs GPT-5.2	–11.6%	[–20.3, –3.0]	–0.30	0.0091**	81
Top-Rated H. vs Claude Opus 4.5	–4.9%	[–13.0, +3.2]	–0.13	0.2300	81
Top-Rated H. vs Gemini 3.0 Pro	–2.0%	[–10.6, +6.7]	–0.05	0.6521	82
Lowest-Rated H. vs GPT-5.2	–23.6%	[–32.3, –14.8]	–0.60	< 0.0001***	81
Lowest-Rated H. vs Claude Opus 4.5	–16.5%	[–25.3, –7.6]	–0.41	0.0004***	81
Lowest-Rated H. vs Gemini 3.0 Pro	–14.1%	[–23.4, –4.7]	–0.33	0.0038**	82
GPT-5.2 vs Claude Opus 4.5	+8.0%	[+0.8, +15.3]	+0.25	0.0294*	80
GPT-5.2 vs Gemini 3.0 Pro	+10.4%	[+4.1, +16.8]	+0.36	0.0016**	81
Claude Opus 4.5 vs Gemini 3.0 Pro	+2.2%	[–2.8, +7.3]	+0.10	0.3784	81

Table 2: **GPT-5.2 exceeds the top-rated human on aggregate review-item quality, and all three AI reviewers exceed the lowest-rated human. Top:** per-group fully-positive rates aggregated paper-level (each paper weighted equally; bootstrap 95% CI) and item-level (papers weighted by item count; Wilson 95% CI). **Bottom:** all 10 paired comparisons on per-paper rates, with paired difference (row – column), 95% CI, Cohen’s *d*, and *p*-value; positive = first-named reviewer higher. **p* < 0.05, ***p* < 0.01, ****p* < 0.001.

across 492 reviews. Second, they identify which of the three human referees wrote the best review and which wrote the worst, giving a top-rated and a lowest-rated human baseline for each paper. Third, for each AI review they indicate which of the three human reviews it matches or exceeds in overall quality, selecting all that apply.

Metrics Correctness and evidence are reported as the percentage of items rated positively within their respective denominators, and significance as the mean score on the 0 to 2 scale. Because the dimensions are applied in sequence, each is computed over a progressively narrower subset: significance over correct items and evidence over correct, at least marginally significant items. An item is *fully positive* when it is correct, at least marginally significant, and sufficiently evidenced. The fully-positive rate is therefore the fraction of a reviewer’s items an author can act on directly. The unit of analysis throughout is the paper ($n = 82$): this makes comparisons invariant to the number of items per review (further results of item-level rates and Generalized Linear Mixed Model analysis are in [Appendix B](#)). Note that GPT-5.2 and Claude Opus 4.5 each declined to review one biology paper, citing safety concerns; those reviews are excluded from the affected comparisons, which is why paired sample sizes vary between 80 and 82. Statistical tests, effect sizes, and confidence intervals follow paired procedures detailed in § 6.

Per-dimension comparison [Table 1](#) reports paper-level means on each dimension together with pairwise effect sizes and *p*-values against each human baseline. On *correctness*, all three AI reviewers fall below the Top-Rated Human (92.3%) by 6 to 10 percentage points: GPT-5.2 reaches 86.2% ($d = -0.23$, $p = .046$), Claude Opus 4.5 reaches 83.7% ($d = -0.34$, $p = .003$), and Gemini 3.0 Pro reaches 81.9% ($d = -0.42$, $p < .001$). On *significance*, however, the direction reverses: among correct items, all three AI reviewers raise more significant criticisms than the Top-Rated Human (mean significance score 1.39 on the 0 to 2 scale), with rank-biserial correlations of $r = +0.49$ for GPT-5.2, $r = +0.30$ for Claude Opus 4.5, and $r = +0.42$ for Gemini 3.0 Pro (all $p \leq .028$). On *sufficiency of evidence*, GPT-5.2 and Claude Opus 4.5 score slightly higher than the Top-Rated Human ($d = +0.23$ and $+0.24$ respectively, both $p < .05$), while Gemini 3.0 Pro is statistically indistinguishable ($d = -0.10$, $p = .380$). Compared to the Lowest-Rated Human, no AI reviewer falls significantly below on any dimension, and all three are significantly higher on significance ($r = +0.43$ to $+0.56$). Together, the results reveal a *tradeoff between correctness and significance*: AI reviewers raise more significant issues, but with lower correctness.

Aggregate review-item quality The dimension-level results above do not directly tell us what fraction of each reviewer’s items would be fully useful to an author: a criticism that is correct but not significant, or significant but insufficiently evidenced, provides little actionable feedback. We therefore examine the *fully positive* rate, defined above. As shown in [Table 2](#), the paper-level mean fully-positive rate is 48.2% for the Top-Rated Human, 36.2% for the Lowest-Rated Human, 60.0% for GPT-5.2, 53.1% for Claude Opus 4.5, and 50.2% for Gemini 3.0 Pro. GPT-5.2 exceeds the Top-Rated Human by 11.6 percentage points ($d = -0.30$, $p = .009$; paired difference Top-Rated – GPT-5.2), while Claude Opus 4.5 and Gemini 3.0 Pro are statistically indistinguishable from the Top-Rated

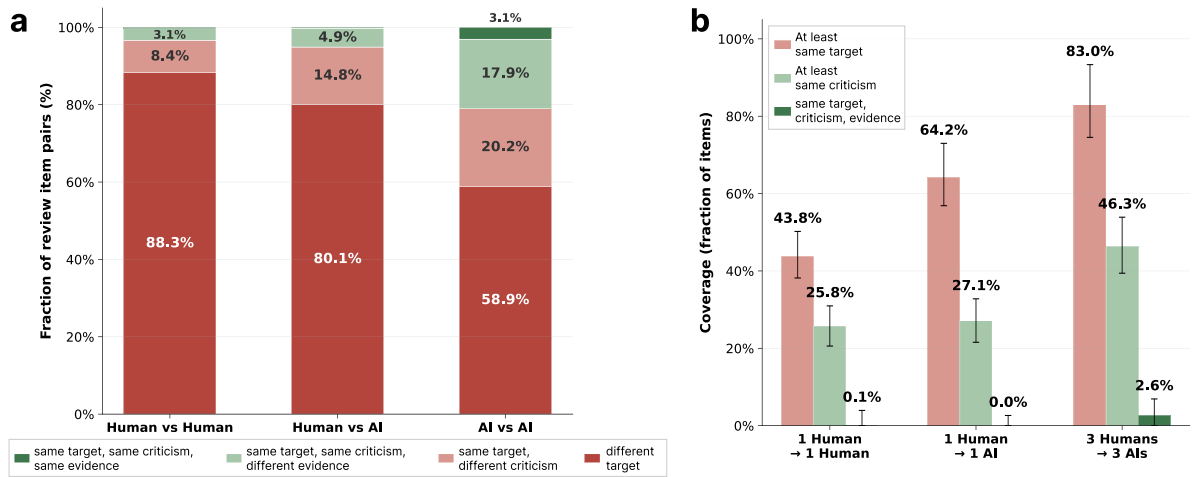


Figure 2: AI reviewers overlap with each other much more than humans do, while AI panels match most human targets but only about half of the specific criticisms. **a**, Distribution of cross-reviewer item pairs across the four similarity categories, for human–human, human–AI, and AI–AI pair types. **b**, Fraction of one reviewer’s items covered by another at three progressively stricter similarity thresholds: at least same target, at least same criticism, exact same evidence. The three configurations compare another human (reference), one AI reviewer, and the union of 3 AI reviewers. Percentages are Rogan–Gladen-corrected; error bars show 95% cluster-bootstrap CIs. Human ($p = .23$ and $p = .65$ respectively). All three AI reviewers substantially exceed the Lowest-Rated Human (differences of +14.1 to +23.6 percentage points, all $p \leq .004$). Among the AI reviewers, GPT-5.2 produces the highest-quality items on average, significantly above both Claude Opus 4.5 ($d = +0.25$, $p = .029$) and Gemini 3.0 Pro ($d = +0.36$, $p = .002$), which in turn are statistically indistinguishable from each other.

Expert-judged win rates Beyond the item-by-item comparison, each domain scientist also provided a paper-level judgment of which AI reviewers matched or exceeded the overall quality of each human reference. As shown in Table 3, these holistic assessments closely track the aggregate quality results. GPT-5.2 is judged to match or exceed the Top-Rated Human on 48.6% of papers and the Lowest-Rated Human on 73.4%. Claude Opus 4.5 and Gemini 3.0 Pro trail GPT-5.2 substantially on the top-rated comparison (32.1% and 30.3%), but both clear the lowest-rated bar on a majority of papers (68.8% and 59.6%). Also, the ordering in Table 3 is consistent with the fully-positive-rate ordering in Table 2, suggesting that expert holistic judgments are well-aligned with the aggregate of per-item ratings.

Reviewer	vs Top-Rated Win-rate [95% CI]	vs Lowest-Rated Win-rate [95% CI]
GPT-5.2	48.6% [38.7, 58.5]	73.4% [64.2, 82.4]
Claude Opus 4.5	32.1% [22.5, 42.0]	68.8% [58.9, 78.4]
Gemini 3.0 Pro	30.3% [21.6, 39.4]	59.6% [50.4, 69.2]

Table 3: Expert scientists judge GPT-5.2 reviews to match or exceed the top-rated human reviewer’s review on nearly half of all papers. Fraction of papers where each AI reviewer matches or exceeds the human reference ($n = 109$ observations across 82 papers, 95% cluster-bootstrap CIs).

3 To what extent do AI reviews overlap with human reviews?

Overview Whereas § 2 compared AI and human reviewers on the quality of each individual review item, this section asks how the criticisms they raise relate to one another, without positioning either source as ground truth. The motivation is that multi-reviewer peer review draws much of its value from the *diversity of perspectives* that different reviewers bring to a manuscript [Page, 2008]: assigning a slate of reviewers with diverse expertise both improves coverage of the manuscript and reduces redundancy across reviews [Goyal et al., 2024]. Whether AI reviewers contribute to or erode this diversity depends not on their individual quality but on whether they raise concerns humans already raise. We measure this by comparing every pair of review items written about the same paper, asking whether the two items point at the same part of the paper (*i.e.*, target), whether they say the same thing about it (*i.e.*, criticism), and whether they cite the same evidence in support. Applying this comparison within and across reviewer types tells us how much two human reviewers duplicate each other, how much an AI reviewer duplicates a human reviewer, and how much of a human panel’s feedback an AI panel would reproduce.

Overlap taxonomy Each review item can be decomposed into three components. The **target** is the specific part of the paper being pointed at: a section, figure, equation, code function, supplementary file, or specific claim. The **criticism** is what the reviewer says about the target: what is wrong with it, what is missing, or what needs

Metric	Value	95% CI
Human items covered by 1 AI reviewer	26.9%	[21.3, 32.6]
Human items covered by 3 AI reviewers	46.3%	[39.4, 53.9]
Fully-positive human items covered by 1 AI	36.3%	[29.6, 43.6]
Fully-positive human items covered by 3 AI	59.2%	[50.2, 69.1]
AI items covered by ≥ 1 human reviewer	74.0%	[65.5, 84.1]
Uncovered AI items (no human match)	26.0%	[15.9, 34.5]
Uncovered AI items that are fully positive	48.1%	[39.6, 56.6]
Uncovered AI items that are correct	81.8%	[75.3, 87.9]

Table 4: **Coverage of human concerns by AI reviewers.** “Covered” = at least one cross-reviewer item pair is classified as similar (same target and same criticism). Per-reviewer-pair averaging at paper level; Rogan–Gladden-corrected.

Metric	Uncovered	Matched	χ^2	p-value
Correctness	81.8% (79)	84.4% (81)	1.45	0.229
Significance $\geq 1^\dagger$	87.4% (75)	91.8% (80)	2.81	0.094
Significance $= 2^\dagger$	57.6% (75)	63.5% (80)	14.98	$< 0.001^{***}$
Evidence sufficient ‡	93.5% (71)	93.2% (80)	0.00	0.971
Fully positive (composite)	48.1% (79)	54.8% (81)	13.71	$< 0.001^{***}$

Table 5: **Uncovered AI items are equally correct and well-evidenced but less often rated as highly significant.** “Uncovered”/“Matched” = AI items with/without a similar human counterpart. N (number of papers) in parentheses. † Among correct items (cascading filter). ‡ Among correct and at least marginally significant items. $***p < 0.001$ (χ^2 tests).

improvement. The **evidence** is what supports the criticism: quoted text from the manuscript, code blocks, or external references. Comparing two review items along these three components yields four mutually exclusive categories: (1) *different target*; (2) *same target, different criticism*; (3) *same target, same criticism, different evidence*; and (4) *same target, same criticism, same evidence* (near-paraphrases). For example, two reviewers who both flag Figure 2 but object to missing error bars and an inaccessible color scheme respectively fall in category (2); if both object to the missing error bars but cite different sentences as evidence, they fall in category (3). We classify a pair as *similar* if it shares the same target and the same criticism (categories 3 and 4), and *not similar* otherwise. The taxonomy was derived from how domain scientists treated overlap in the free-form paper-level survey responses (see § 6.6).

Analysis setup We compare three strata of within-paper reviewer pairs: human–human (H–H), human–AI (H–A), and AI–AI (A–A), under two complementary views of the same 65,704 pairs. The **pair-level** view (Figure 2a) gives the distribution of pairs over the four categories; the **coverage** view (Figure 2b; Table 4) instead takes each review item as the denominator and asks whether the comparison reviewer supplies *any* counterpart meeting a given threshold (at least same target, at least same criticism, exact same evidence). The two views differ in denominator and are not directly comparable. Similarity is scored by an LLM judge (GPT-5.4; 92.7% binary accuracy, sensitivity 87.1%, specificity 96.8%) and rates are Rogan–Gladden-corrected [Rogan and Gladen, 1978] within stratum, computed per paper, and reported with 95% cluster-bootstrap CIs (10,000 paper-level resamples). Comparisons between strata are read from these intervals; component-level comparisons of uncovered versus matched AI items (Table 5) use χ^2 tests. Full details are in § 6.6.

Human reviewers barely overlap The H–H baseline in Figure 2a establishes what normal overlap looks like, and it is strikingly low. Only 3.4% of pairs drawn from two different human reviewers of the same paper share both a target and a criticism; the remaining 96.6% either raise different criticisms about the same target (8.4%) or address entirely different targets (88.3%). The coverage view gives the same picture at the item level: 25.8% of one human reviewer’s items have a same-criticism counterpart in another human’s review, and 43.8% have any same-target counterpart at all. Human peer review in our dataset is therefore not a redundant exercise. Each reviewer brings a largely distinct set of concerns, which provides empirical support for the *diversity of perspectives* argument for peer review [Page, 2008] and supplies the reference point against which the AI comparisons below should be read.

AI reviewers raise valid concerns humans miss A natural next question is whether AI reviewers add anything human reviewers do not, and whether those additions are valuable contributions or spurious artifacts. Table 4 shows that 74.0% of AI-raised items have a similar counterpart in at least one human’s review (*matched*), leaving 26.0% with no similar human counterpart (*uncovered*). A component-level comparison of uncovered versus matched AI items (Table 5) shows that uncovered items are not hallucinations and are not under-evidenced: their correctness rate (81.8%) and evidence-sufficiency rate (93.5%) are both statistically indistinguishable from those of matched AI items (84.4% correct, $p = 0.23$; 93.2% evidence-sufficient, $p = 0.97$). What separates the two groups is the fraction rated as *highly* significant: 57.6% for uncovered versus 63.5% for matched ($p < 0.001$), with the same gap appearing across all three AI reviewer models. This does not contradict the § 2 finding that AI reviewers as a whole raise more significant items than human reviewers: AI items that overlap with human concerns tend to be the items humans also flagged, while uncovered AI items, by definition, correspond to observations humans did not raise. We interpret this as AI reviewers reliably surfacing the major concerns human reviewers would also raise, and *additionally* surfacing a non-trivial volume of valid, well-supported observations that humans miss.

Replacing one human with AI doesn’t hurt diversity Substituting a single AI reviewer for a single human leaves the pairwise diversity of a panel nearly intact. The rate at which an AI reviewer overlaps with a human reviewer

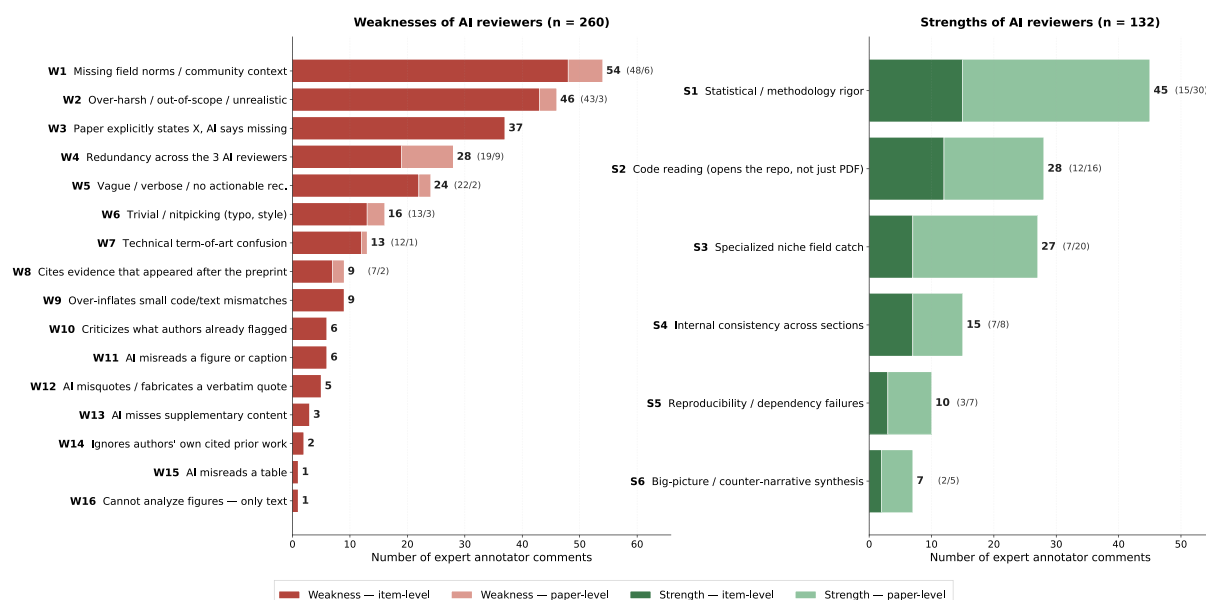


Figure 3: **Strengths and weaknesses of AI reviewers identified by domain experts.** Distribution of 442 free-form comments on AI reviews across 16 weakness categories (left, $n = 260$) and 6 strength categories (right, $n = 132$). Dark bars are item-level comments; light bars are paper-level comments. Categories are sorted by total count.

(H–A: 5.1% [0.3, 9.0]) is only slightly above the 3.4% human–human baseline, and the confidence intervals of the two estimates overlap substantially (Figure 2a). The coverage view gives the same ranking at the item level: the fraction of one human reviewer’s items with a same-criticism counterpart rises only marginally when the second reviewer is an AI rather than another human (26.9% versus 25.8%). An AI reviewer added to a panel therefore behaves, in terms of what it duplicates, much as an additional human reviewer would.

AI panels diverge from human panels The same does not hold when the entire panel is replaced. AI reviewers overlap with one another at 20.9% [16.2, 25.4] on target and criticism (Figure 2a), roughly six times the human–human baseline, so a panel of three AI reviewers is substantially redundant with itself. Figure 2b shows the consequence for coverage: the union of items raised by three AI reviewers covers 83.0% of the human-raised items at the same-target level but only 46.3% at the same-criticism level, and a single AI reviewer already reaches 64.2% same-target coverage against 43.8% for a second human. For authors, this means that replacing three human reviewers with three AI reviewers would return feedback about most of the same parts of their paper but only about half of the specific criticisms with the same substance. AI reviewers and human reviewers converge on which parts of a paper warrant review while diverging in how they characterize what they find, which makes an AI panel a poor drop-in replacement for a human panel even though a single AI reviewer is a reasonable addition to one.

4 What are the concrete strengths and weaknesses of AI reviewers?

Overview § 2 and § 3 characterize how AI reviewers compare to human reviewers, but they do not describe the specific behaviors that drive those differences. For researchers studying AI review systems, editors considering AI-assisted review, and authors receiving AI-assisted feedback alike, a more pointed question is which behaviors of current AI reviewers most undermine their usefulness, and which genuinely outperform human reviewers. We answer this from the free-form comments annotators wrote alongside their ratings, reading every comment that names a concrete strength or weakness of an AI review relative to a human one and grouping them into recurring categories. The resulting picture is consistent across all three AI reviewers: their weaknesses are failures of context, and their strengths are forms of scrutiny that reward diligence.

Analysis setup Annotators produced 767 item-level and 250 paper-level comments across the 82 papers (see Figure 9). Of these, 392 name a concrete strength or weakness of AI reviewers relative to human reviewers. We manually categorized them into 16 weakness ($n = 260$) and 6 strength ($n = 132$) categories, iterating on the category definitions until two authors could apply them independently; Figure 3 gives the full distribution.

Strengths of AI reviewers The four most frequently cited strengths account for 87% of strength comments (115/132), and they describe a consistent pattern: AI reviewers are more diligent than human reviewers at checking that requires effort rather than judgment. Statistical and methodological rigor (S1, $n = 45$) is the largest category and appears across all three models: annotators credited AI reviewers with catching independence violations between

training and test splits, absent uncertainty estimates on headline numbers, and sample sizes too small to support the claimed effect. Source code inspection (**S2**, $n = 28$) reflects the AI reviewers' filesystem access to everything the authors submitted: they opened the repository alongside the manuscript and supplementary files, and found implementation bugs and data leakage that the text alone does not reveal. Subfield technical depth (**S3**, $n = 27$) appears where a method or term carries commitments the paper does not discharge, such as invoking a technique whose stated assumptions the reported setup violates. Internal consistency (**S4**, $n = 15$) covers mismatches between what the abstract claims, what the methods describe, and what the results report.

Limitations of AI reviewers The five most frequently cited weaknesses account for roughly three quarters of weakness comments (189/260), and they describe AI reviews as poorly calibrated to context rather than factually empty. The largest theme is severity miscalibration (**W1** missing community or field norms, $n = 54$; **W2** over-harsh or out-of-scope demands, $n = 46$): the AI demands a validation the field does not require, or treats a standard practice as a flaw, because it applies a discipline-neutral standard for reproducibility, sample size, or out-of-distribution testing without knowing what the subfield treats as routine. The content is usually defensible in the abstract; the weight placed on it is not. The second theme is reading failure (**W3** paper explicitly states X , AI says missing, $n = 37$): the AI asserts that a detail is absent when the paper supplies it in a different section, in the supplementary material, or a few sentences later in the same passage. The third theme concerns the review as an artifact (**W4** redundancy across the three AI reviewers, $n = 28$; **W5** vague or verbose reviews without actionable recommendations, $n = 24$): the three AI reviewers raise overlapping criticisms, consistent with the AI–AI overlap in § 3, and individual reviews run long without arriving at a concrete revision request.

In § D.1, we present 24 detailed case-study examples spanning the nine most informative categories, each shown with the original review item, the expert annotator's commentary, and our discussion.

5 Conclusion

In this paper, we characterize the strengths and weaknesses of AI reviewers through an expert annotation study in which 45 domain scientists, over 469 hours, evaluate all 2,960 review items from both AI and human reviewers across 82 *Nature*-family papers. Current AI reviewers are competitive even against the strongest reviewer a paper actually received: on the composite of correctness, significance, and evidence sufficiency, GPT-5.2 produces a higher fraction of fully positive items than the top-rated human referee, and all three models exceed the lowest-rated one. This comes at a cost in correctness, offset by higher significance among the items they get right. Their failures concentrate in misjudged severity against subfield conventions, lost track of content across long papers, and inflated minor issues, while their strengths lie in scrutiny that rewards diligence. AI reviewers also duplicate one another far more than human referees do, so the panel and the individual behave differently.

Looking ahead, three implications follow. For developers, correctness is the bottleneck: AI reviewers already find substantive issues, and the failures we observe are failures of accumulated field knowledge, so the tractable interventions are giving reviewers access to community norms and better long-context management. For editors and venues, the overlap results argue that AI reviewers belong in a panel rather than as a panel: adding one preserves the diversity a multi-reviewer process depends on, while replacing three human referees with three AI reviewers returns most of the same targets but half the substance. For the field, AI reviews should not be evaluated against human reviews as an implicit gold standard, but on the same per-dimension terms domain experts apply to human reviews, since two reviews reaching identical verdicts can share almost no substance. Open questions remain, including whether the correctness–significance tradeoff persists as models improve, whether these patterns hold outside *Nature*-family papers, and what governance norms should accompany operational deployment. The framework and annotations we release provide the empirical infrastructure to answer them, and the urgency only grows as AI reviewers move further into deployment.

An example of review item # 1 from a human review

"In Fig. 2, the appropriate comparison could include Carr-Parrinello MD since that is common to use for minor distortion of the geometry. At a minimum, mentioning CPMD in the main text would be less misleading to the reader."

An example of review item # 2 from a human review

"Would the author advise using CASSCF, RASSCF, or better CASPT2 to properly include both static and dynamic correlations? That could represent an important application for the method, especially in the framework of dynamics. CAS-based dynamics is highly employed and reliable, being able to accelerate it by reducing the number of electronic structure calls using the approach proposed would be an important achievement."

Figure 4: Two example review items written by the same human reviewer of a paper in the Physical Sciences. Each **review item** corresponds to a single atomic criticism directed at one aspect of the paper, making up the entire review.

6 Methods

6.1 The criticism-level evaluation framework

Unit of analysis Our framework compares reviews at the level of the individual criticism rather than the verdict. We define a **review item** as a single atomic criticism directed at one specific aspect of the paper, and use this term throughout for the unit the framework operates on. This contrasts with prior evaluations of AI reviewers, which compare reviews at the aggregate level (e.g., overall score, accept/reject verdict). A single peer review typically contains multiple distinct criticisms, and although separating free text into atomic criticisms is non-trivial in general, human reviewers conventionally use bullet points, enumeration markers (e.g., "First," "Second,"), or paragraph transitions to demarcate the points they want authors to address in revision. We rely on these markers to manually decompose each human peer review into review items; AI reviewers produce items in structured form (§ 6.3). Figure 4 shows two example review items extracted from the same human peer review of a physical-sciences paper.

Adjudication dimensions Domain scientists with subfield-matched expertise rate every review item in their assigned paper along three dimensions:

- **Correctness** (binary): whether the main point of the criticism is correct (i.e., the issue it raises actually exists in the paper rather than being a misreading of the manuscript) and is clearly stated.
- **Significance** (ordinal, three-level): conditional on the criticism being correct, whether it addresses a significant aspect of the paper. The three levels are *Significant* (an insightful concern that, if addressed, would meaningfully improve the paper), *Marginally Significant* (e.g., typos or stylistic issues), and *Not Significant* (a minor item that would be better removed from the review).
- **Sufficiency of evidence** (binary): conditional on the criticism being correct and at least marginally significant, whether the evidence accompanying the criticism (e.g., quotes from the paper's main text, supplementary materials, or external references) is sufficient to support the main point.

The **cascading structure** reflects the logical dependency among the three dimensions: significance is only meaningful for criticisms that are correct, and evidence sufficiency is only meaningful for criticisms that are correct and at least marginally significant. We chose a three-dimensional design rather than a single overall rating so that we can identify the specific aspects in which AI reviewers are better or worse than human reviewers. Correctness and evidence sufficiency are reported as the percentage of items rated positively within their respective denominators, and significance as the mean score on a 0 to 2 scale.

Composite indicator Alongside the three dimensions we report **fully positive**, an integrative indicator that an item is correct, at least marginally significant, and sufficiently evidenced. This is the fraction of a reviewer's items an author could act on without first having to discount them, and it is the primary aggregate quality measure in § 2. Because the dimensions cascade, an item that fails correctness fails the composite regardless of its unrated downstream dimensions.

Dual human baseline Referees of the same paper vary widely in quality, so comparing AI reviews against a single averaged human reference would obscure whether an AI reviewer competes with the best review a paper received or merely with the weakest. The framework therefore retains two human baselines per paper, the top-rated and the lowest-rated referee, identified by the annotator in the paper-level survey (§ 6.4).

6.2 Paper selection and dataset

Selection criteria We chose 82 papers from *Nature* and its sister journals, spanning the physical, biological, and health sciences. Papers are included only if they meet three criteria: (1) a publicly released set of official human peer reviews under *Nature*’s transparent peer review policy, so that we have human reviews for comparison; (2) a publicly available pre-review version of the manuscript on Research Square¹, so that AI reviewers evaluate the same manuscript that the human reviewers did; and (3) a subfield match with one of our recruited domain scientists, so that every review item can be annotated by an expert with relevant methodological knowledge. These three constraints jointly narrow the candidate pool substantially: public peer review is not the default at most venues, pre-review versions are rarely available after publication, and subfield-specific expert recruitment further restricts viable subjects.

Dataset statistics The 82 papers are drawn from *Nature Communications* (73 papers), *Nature* (2), *Nature Computational Science* (2), *Nature Ecology & Evolution* (2), *Nature Methods* (1), *Nature Physics* (1), and *Nature Microbiology* (1)², published between 10 January 2020 and 27 October 2025. Following the *Nature Communications* subject taxonomy³, the 82 papers span 27 subject categories: 38 in Physical Sciences, 30 in Biological Sciences, and 14 in Health Sciences. Beyond the main manuscript text, most submissions include additional components that AI reviewers can also access: 83% have supplementary materials, 76% have separately submitted figures, and 74% have submitted source code. Per-paper content and review statistics are summarized in Table 6.

Component	Mean	Median	Min	Max
<i>Manuscript content</i>				
# Words in main manuscript	8,820	8,500	3,787	22,830
# Supplementary files per paper	5.5	2	0	129
# Figures per paper	4.0	4	0	19
# Code files per paper	296.7	18	0	8,285
<i>Peer review</i>				
# Human reviews per paper	2.9	3	1	5
# Review items per human review	8.8	7	1	72
# Review items per AI review	4.2	4	2	5

Table 6: Statistics of the 82 papers in the expert annotation study.

6.3 Reviewers: official human referees and frontier-LLM agents

Human reviewers For each paper, we use the first-round official human peer reviews released by *Nature*-family journals, retaining the first three reviewers when more than three are present. Each review is decomposed into review items per § 6.1. Further details are in § A.4.

AI reviewers We use three frontier language models as AI reviewers: GPT-5.2, Claude Opus 4.5, and Gemini 3.0 Pro. Each model is deployed as an agent through the OpenHands software-agent-sdk [Wang et al., 2026], with filesystem access to the paper’s source files (main text, supplementary materials, figures, and submitted code) and a small set of tools (shell terminal, file editor, task tracker, and a web-search tool with the paper’s publisher domains blocked to prevent retrieval of the published version or peer review report). Each agent receives a prompt asking it to produce up to five review items per paper according to the six *Nature* peer-review evaluation criteria, with each item structured as a **main claim** (the central point of criticism, with its associated criterion) followed by **supporting evidence** (a set of quotes from the paper’s main text, supplementary materials, submitted source code, or external references, each accompanied by an interpretive comment). Figure 5 shows an example AI-produced review item for the same paper as Figure 4. We generated one review per (paper, model) pair. GPT-5.2 and Claude Opus 4.5 each declined to review one biology paper, citing safety concerns that do not apply to the manuscripts in question; no substitute review was generated, and those papers are excluded from the affected comparisons.

6.4 Expert annotators

Recruitment Our annotator pool comprises 45 domain scientists from 25 institutions: 23 faculty members, 7 research scientists at industrial labs, national laboratories, or research institutes, 6 postdoctoral researchers, and 9 Ph.D. students. They produced 109 sets of annotations across the 82 papers (averaging 2.42 papers per scientist), totaling 469 hours of expert annotation. Further details are in § A.7.

Annotation task Each annotator receives one assigned paper together with its six reviews, three written by human referees and three generated by AI reviewers, and provides three kinds of annotation. First, they rate every review item from all six reviews on the three dimensions in § 6.1, with each rating accompanied by a free-form comment justifying the judgment. Second, they complete a paper-level survey selecting the **Top-Rated Human Reviewer** and **Lowest-Rated Human Reviewer** from the paper’s official referees based on overall review quality, which supplies

¹<https://www.researchsquare.com/>

²All manuscripts and peer review files in our dataset are released by their publishers under a CC BY 4.0 license.

³<https://www.nature.com/ncomms/browse-subjects>

An example of review item #1 from an AI review

Claim: “The generalized eigenproblem is not made numerically robust”

- **Main point of criticism:** The paper’s implementation of the core generalized eigenvalue solve is not numerically hardened against overlap-matrix ill conditioning, so the reported active-learning MD workflow can fail unpredictably once training states become nearly linearly dependent.
- **Evaluation criteria:** Data and methodology; Validity

Evidence:

Quote 1 (from main text): “This is achieved in closed form as the diagonalization of a generalized eigenvalue problem in the basis of the training states.”

| **Comment:** This is the numerical kernel of the method, so its stability determines whether the algorithm can actually be used in the long active-learning trajectories described in the paper. Because the training basis is enlarged sequentially, the overlap matrix can become nearly singular as similar geometries are added. The manuscript does not describe any canonical orthogonalization, eigenvalue truncation, or other conditioning step for this solve. That omission is a material reproducibility risk, not a minor implementation detail. A paper that claims black-box molecular dynamics should show how this step remains stable as the basis grows.

Quote 2 (from submitted source code):

```
if hermitian is True:
    # Solve the generalized eigenvalue problem for Hermitian
    # Hamiltonian
    vals, vecs = eigh(H, S)
```

| **Comment:** The implementation delegates the solve directly to `scipy.linalg.eigh` with the overlap matrix passed in unchanged. There is no diagnostic for the condition number of `S`, no thresholding of small eigenvalues, and no fallback path if the basis is nearly linearly dependent. A minimal reproduction of this code on a singular overlap matrix raises an exception, which shows the failure mode is real rather than hypothetical. Because the MD loop depends on repeating this operation many times, the current implementation is too brittle for the robustness claims made in the manuscript. The code needs a conditioning strategy before the method can be treated as reliably black-box.

Quote 3 (from external reference, SciPy documentation): “If eigenvalue computation does not converge, an error occurred, or `b` matrix is not definite positive. Note that if input matrices are not symmetric or Hermitian, no error will be reported but results will be wrong.”

| **Comment:** This warning from SciPy documents the failure mode that the current code is exposed to when the overlap matrix becomes singular or nearly singular. The paper does not report any safeguards to ensure that every reported run stays safely away from that regime. A direct reproduction of the current solver on a singular overlap matrix triggers exactly this kind of `LinAlgError`, showing that the instability is not merely theoretical. To support its robustness claims, the manuscript should either regularize the overlap matrix or report the conditioning of `S` for the presented trajectories. Without that, the method’s reliability remains unproven.

Figure 5: An example of a review item produced by an AI reviewer for the same paper as Figure 4. A review item consists of a **Claim** (main point of criticism and evaluation criteria) and **Evidence** (a set of supporting quotes with interpretive comments). Evidence quotes may be drawn from the paper’s main text, supplementary materials, the source code, or external references. We instruct AI reviewers to produce at most five review items per paper.

the dual baseline used in § 2 and § 3. Third, they indicate for each AI reviewer which of the human references it matches or exceeds in overall quality, and optionally note criticisms that one reviewer raised which the others missed. The qualitative responses are analyzed in § 4; the free-text notes on missed criticisms also motivated the similarity taxonomy in § 6.6.

358
359
360
361

Inter-annotator agreement To measure inter-annotator agreement, 27 of the 82 papers were independently annotated by a second domain scientist, yielding 908 doubly-annotated review items. Because the marginal class distributions of our annotations are highly skewed, we report Gwet’s AC1 alongside raw percent agreement and Cohen’s κ in Table 7. Agreement is almost perfect for the two binary dimensions (correctness and evidence sufficiency) and moderate for the three-level significance scale; the substantial gap between κ and AC1 on the two skewed binary dimensions illustrates the well-known kappa paradox under skewed marginals, motivating our choice to report AC1 as the primary chance-corrected measure [Feinstein and Cicchetti, 1990, Gwet, 2008]. The full IRR analysis is in § A.9.

Dimension	N	% Agree.	Cohen’s κ	Gwet’s AC1
Correctness (binary)	908	85.8%	0.28 (fair)	0.82 (alm. perf.)
Significance (3-way ord.)	743	59.9%	0.31 (fair)	0.44 (moderate)
Evidence (binary)	583	88.0%	0.12 (slight)	0.86 (alm. perf.)

Table 7: Inter-annotator agreement on 908 doubly-annotated review items from 27 doubly-annotated papers. N decreases across dimensions due to the cascading annotation structure. Interpretation labels follow Landis and Koch [1977].

6.5 Statistical analysis

Unit of analysis Throughout § 2 and § 3 the unit of analysis is the paper ($n = 82$) rather than the review item. Item counts differ across reviewers because AI reviews are capped at five items per paper while human reviews are not; aggregating to paper-level rates before testing makes comparisons invariant to the number of items per review. For each paper we pool item-level annotations: items from the 27 doubly-annotated papers contribute two annotation rows each, while items from the remaining 55 contribute one. This respects within-paper correlation among items rated by the same domain scientist. The expert-judged win rates in Table 3 are the exception, treating each of the 109 annotation sets as a separate observation with cluster-bootstrap intervals over papers, because the paper-level survey is completed once per annotator rather than once per paper.

Tests and effect sizes Binary dimensions and the fully-positive composite are compared with two-sided paired t -tests on paper-level rates, reported with Cohen’s d_z . The ordinal significance score is compared with the two-sided Wilcoxon signed-rank test, reported with rank-biserial correlation r . All estimates carry 95% bootstrap confidence intervals from 10,000 paper-level resamples (percentile method). We report unadjusted p -values throughout. Papers for which a model produced no review are excluded pairwise rather than imputed, which is why paired sample sizes range from 80 to 82. Item-level rates and a generalized linear mixed model with paper-level random intercepts reach the same conclusions (Appendix B).

6.6 Measuring overlap between review items

Provenance of the similarity taxonomy The target/criticism/evidence decomposition used in § 3 was derived inductively from the free-form paper-level survey responses described in § 6.4, in which annotators noted criticisms that one reviewer raised and others missed. Scientists consistently treated two items as “the same concern” when they pointed at the same part of the manuscript and made the same substantive objection, irrespective of whether the two reviewers cited the same supporting material. The four-category scheme formalizes that distinction. Additional examples and the coding procedure are in § C.2.

Two views of the same pairs We report overlap under two views. The **pair-level** view treats each cross-reviewer item pair as a unit and gives the distribution of pairs over the four mutually exclusive categories. The **coverage** view treats each review item as a unit and asks whether the comparison reviewer supplies at least one counterpart meeting a given threshold, where the thresholds are cumulative unions of the categories: *at least same target* corresponds to categories 2–4, *at least same criticism* to categories 3–4, and *exact same evidence* to category 4 alone. The two views answer different questions and use different denominators, so their percentages are not directly comparable: a low pair-level rate is compatible with substantially higher coverage, because a single item may be matched by any one of the comparison reviewer’s items.

Similarity judge Exhaustive manual annotation of all 65,704 within-paper cross-reviewer pairs is infeasible, so we score similarity with an LLM judge. The judge receives both review items in full (claim and evidence) together with the four-category definition, and returns a single category. We calibrated the judge against 164 pairs annotated manually, on which the selected model (GPT-5.4) achieves 92.7% binary accuracy (similar vs. not similar) and 83.5% accuracy on the full four-way classification, with sensitivity 87.1% and specificity 96.8%. The comparison of candidate judges and the full confusion matrix are in Appendix C.

Prevalence correction A judge with imperfect sensitivity and specificity yields a biased estimate of the true similarity rate, and the direction of that bias depends on the underlying rate itself. We therefore convert each

apparent rate p_{obs} into a corrected estimate using the Rogan–Gladden formula [Rogan and Gladden, 1978], $p_{\text{true}} = (p_{\text{obs}} + \text{spec} - 1) / (\text{sens} + \text{spec} - 1)$, with estimates truncated to $[0, 1]$. The correction is applied within each stratum, which assumes the judge’s sensitivity and specificity are constant across pair types. The correction applies to similarity and coverage rates, which are prevalence estimates. It does not apply to the quality rates of uncovered items reported in Table 4 and Table 5 (e.g., the fraction of uncovered AI items that are correct), which are conditional on expert annotations rather than judge output; we note, however, that membership in the uncovered set is itself determined by the judge, so judge error propagates into these conditional rates even though the rates are not corrected.

Aggregation and inference All overlap statistics use the paper as the unit of analysis, consistent with § 6.5. Within each paper, rates are computed per reviewer pair and averaged over pairs before averaging across papers, so that papers with more human reviewers do not receive greater weight. Confidence intervals are 95% percentile intervals from a 10,000-iteration cluster bootstrap that resamples papers with replacement and, on each iteration, redraws the judge’s sensitivity and specificity from their sampling distribution, so that judge uncertainty and cross-paper variance are propagated jointly. Comparisons between strata (H–H, H–A, A–A) are read from these intervals rather than from a separate significance test. Comparisons of uncovered versus matched AI items on each annotation dimension use χ^2 tests across papers contributing both categories, with the number of contributing papers reported in Table 5; the denominator varies across rows because of the cascading annotation structure described in § 6.1.

7 Related Work

Peer review system for science. The peer review system, fundamental to scientific publishing, has come under increasing strain in recent years. Manuscript submission volumes continue to rise across both journals and conferences while the pool of qualified human reviewers does not expand at the same rate, leading to reviewer fatigue, longer review timelines, and difficulty recruiting reviewers willing to provide thorough evaluations [Hanson et al., 2024, Adam, 2025, Chen et al., 2025, Bauchner and Rivara, 2024]. As a result, median submission-to-publication times have stretched to 100 to 160 days at major science journals [Powell, 2016]. These growing pressures have raised concerns about review quality and consistency: the NeurIPS 2014 and 2021 consistency experiments, in which roughly 10% of submissions were independently reviewed by two committees, found that around half of the accepted papers would have been rejected if the review process had been independently rerun, and that approximately 26% (in 2014) and 23% (in 2021) of papers received inconsistent accept-or-reject decisions across the two committees [Cortes and Lawrence, 2021, Beygelzimer et al., 2023], indicating that a substantial portion of decisions reflects subjective variation rather than the underlying quality of the work. Beyond consistency, recent work has documented further structural challenges including reviewer miscalibration, reviewer-author collusion, and undetected major errors in submitted manuscripts [Shah, 2022].

In response, the scientific community has begun to consider whether natural language processing and large language models might assist (rather than replace) the human peer-review process, ranging from low-level support tasks such as plagiarism detection and reviewer-paper matching to higher-level tasks such as structured feedback generation and consistency checks across reviews [Kuznetsov et al., 2024]. This consideration is no longer purely theoretical. A recent global survey of more than 1,600 active researchers across 111 countries reported that over 50% of reviewers have already used AI tools while peer-reviewing manuscripts, often without clear policy guidance and in some cases against the official rules of the venue they were reviewing for [Naddaf, 2026]. At the venue level, several conferences have begun to formally pilot AI assistance within their review workflows: NeurIPS 2024 piloted an opt-in LLM-based author checklist assistant on a voluntary subset of submissions [Goldberg et al., 2024], ICLR 2025 ran a large-scale randomized controlled trial of an LLM-based feedback agent that nudged human reviewers toward more specific and actionable comments across more than 20,000 reviews [Thakkar et al., 2025], and AAAI-26 went a step further and deployed a state-of-the-art AI reviewer on every one of its 22,977 main-track submissions, with surveys of authors and program-committee members reporting that participants found the AI reviews useful and, on certain dimensions such as technical accuracy and research suggestions, preferred them to human reviews [Biswas et al., 2026]. Together, these developments mark a shift from speculative discussion of AI in peer review to its concrete operational deployment at scale.

LLM-as-a-Judge. A parallel line of work has investigated the use of large language models themselves as evaluators of model and text outputs (LLM-as-a-Judge), establishing strong correlations with human judgement on general tasks [Zheng et al., 2023, Kim et al., 2023, 2024] while documenting systematic biases such as position, verbosity, and self-preference effects [Ye et al., 2024, Wataoka et al., 2024, Dubois et al., 2024, Kim et al., 2025, Ye et al., 2025].

Development of AI reviewers. Building on LLM-as-a-Judge methodologies, an active line of research has explored applying LLM-as-a-Judge specifically to scientific peer review (*i.e.*, AI Reviewers). Early exploratory studies tested whether GPT-4 could detect inserted errors and produce structured first-pass feedback [Yuan et al., 2022, Liu and Shah, 2023]. Subsequent work has produced a wide range of full review-generation systems, including multi-agent discussion [D’Arcy et al., 2024], standardized review generation through fine-tuning [Yu et al., 2024], bias-aware pipelines [Tyser et al., 2024], deep-thinking review generation [Zhu et al., 2025], and RL-trained reviewers [Zeng et al., 2025]. A separate strand uses LLMs as a quality-control component within end-to-end research-automation systems [Lu et al., 2026], while another simulates the peer-review process itself with LLM agents to study its dynamics [Jin et al., 2024]. Beyond review generation, several efforts target review-process improvement directly, including a large-scale ICLR 2025 randomized study showing that LLM feedback agents can make human reviewers’ reviews more informative [Thakkar et al., 2026], LLM checklist assistance for authors [Goldberg et al., 2024], and pairwise comparison through LLM agents [Zhang et al., 2026].

Notably, most existing evaluations of AI reviewers operate at the review level rather than the individual-comment level: they measure score correlation, decision alignment, or text-level similarity against human reviews and treat human reviews as the gold standard [Idahl and Ahmadi, 2025, Bougie and Watanabe, 2025]. Some evaluation frameworks have begun to introduce finer-grained components: ReviewEval performs per-claim factual verification and per-insight actionability scoring [Garg et al., 2025]; TreeReview evaluates individual feedback comments through embedding-based and LLM-based matching for coverage and specificity [Chang et al., 2025]; and FLAWS benchmarks LLMs’ ability to localize specific errors in scientific manuscripts [Xi et al., 2025]. Other work has begun to question the gold-standard assumption itself, either by demonstrating that LLM pairwise comparison can identify high-impact papers more accurately than rating-based aggregation [Zhang et al., 2026] or by reporting that AI reviewers can match human reviewers in predicting paper outcomes [Bougie and Watanabe, 2025].

In contrast, our work performs an evaluation that, to our knowledge, no prior framework has performed: every individual review item, from both AI and human reviewers, is annotated by a domain scientist in each paper’s own field along three complementary quality axes (correctness, significance, and sufficiency of evidence). Three properties of this design are absent from prior practice taken together. The annotation is bidirectional, in that AI and human review items are subject to the same scrutiny rather than human reviews being treated as the implicit reference for AI [Chang et al., 2025, Garg et al., 2025, Idahl and Ahmadi, 2025, Bougie and Watanabe, 2025]. The annotation is performed by domain scientists rather than by LLM judges or graduate-student annotators from outside the paper’s subfield [Ryu et al., 2025], since judgments of scientific significance and evidence sufficiency are field-dependent. And the annotation is multi-axis rather than single-dimensional or holistic, since a comment can be correct but trivial, significant but unsupported, or both significant and well-supported, and these are very different failure modes. The result is per-axis directional evidence on the specific quality dimensions on which AI reviewers outperform human reviewers and on which the reverse holds.

Data availability

The annotation data supporting this study, including the identifiers of the 82 *Nature*-family papers, the 2,960 annotated review items, and the per-paper scores, are available as the PeerReview Bench dataset at this <https://huggingface.co/datasets/prometheus-eval/peerreview-bench> (released under CC-BY-4.0).

Code availability

Custom code central to the results is available at this <https://github.com/prometheus-eval/cmu-paper-reviewer> (released under Apache 2.0). A free live demo of the cmu paper reviewer is available at cmu-paper-reviewer.com.

Author contributions

S.K. led the project. The study was conceived and designed by S.K., D.Y., K.G., X.Y., A.A., S.W. and G.N. The AI reviewer pipeline was implemented by S.K., who also ran all reviewing experiments, built the annotation interface, performed the statistical analysis and the similarity classification, and developed the CMU Paper Reviewer platform; the OCR pipeline for manuscript ingestion was developed by J.S. Domain expert scientists were recruited by S.K., D.Y. and K.G. Official referee reports were collected from *Nature*-family journals and the corresponding manuscripts from Research Square by D.Y. Free-form expert comments were manually categorized by S.K. and D.Y. J.B., P.A. and I.W. contributed technical discussions from the early stages of the project. Expert annotation of review items was performed by V.Z., S.P., D.R.S., I.D., F.S., B. Mitreska, Y. Jeong, K.K., Y.M.S., D.M., A.P., B. Mojsoska, M.T., M.S., R.L., H.J.L., N.U.D., Y. Jo, S.H., C. Lee, H.L., E.H.R.T., E.S., K.S., Y.C., J.P., A.S., H.C., Y.S., E.K., A.M., J.A., C. Langkammer, H.M., C.Y., N.S., M.J., E.C., J.K., Y.S.J., W.Y.K., J.K.K., I.M.A., H.U.K.

and D.B. Advisory: C. Lawrence, X.Y., A.O., A.A., S.W. and G.N. Manuscript writing was led by S.K. and D.Y.; all authors contributed to editing and approved the final version. Correspondence to S.K., S.W. and G.N.	523 524
Competing interests	525
The authors declare no competing interests.	526
References	527
Frank Gannon. The essential role of peer review. <i>EMBO reports</i> , 2001.	528
Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. Peer review in scientific publications: benefits, critiques, & a survival guide. <i>Ejifcc</i> , 25(3):227, 2014.	529 530
Kyle Siler, Kirby Lee, and Lisa Bero. Measuring the effectiveness of scientific gatekeeping. <i>Proceedings of the National Academy of Sciences</i> , 112(2):360–365, 2015.	531 532
Nuo Chen, Moming Duan, Andre Huikai Lin, Qian Wang, Jiaying Wu, and Bingsheng He. Position: The current ai conference model is unsustainable! diagnosing the crisis of centralized ai conference. <i>arXiv preprint arXiv:2508.04586</i> , 2025.	533 534 535
Kendall Powell. Does it take too long to publish research? <i>Nature</i> , 530(7589):148–151, 2016.	536
Ryan Liu and Nihar B Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing. <i>arXiv preprint arXiv:2306.00622</i> , 2023.	537 538
Iliia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. What can natural language processing do for peer review? <i>arXiv preprint arXiv:2405.06563</i> , 2024.	539 540 541
Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, et al. Ai-assisted peer review at scale: The aaai-26 ai review pilot. <i>arXiv preprint arXiv:2604.13940</i> , 2026.	542 543 544
Arjun K Manrai, David Ouyang, Joseph W Hogan, and Isaac S Kohane. Accelerating science with human+ ai review, 2025.	545 546
Ahmed Saad, Nathan Jenko, Sisith Ariyaratne, Nick Birch, Karthikeyan P Iyengar, Arthur Mark Davies, Raju Vaishya, and Rajesh Botchu. Exploring the potential of chatgpt in the peer review process: an observational study. <i>Diabetes & Metabolic Syndrome: Clinical Research & Reviews</i> , 18(2):102946, 2024.	547 548 549
Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 29330–29355, 2025.	550 551 552
Maximilian Idahl and Zahra Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. In <i>Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)</i> , pages 550–562, 2025.	553 554 555 556
Yaohui Zhang, Haijing ZHANG, Wenlong Ji, Tianyu Hua, Nick Haber, Hancheng Cao, and Weixin Liang. From replication to redesign: Exploring pairwise comparisons for LLM-based peer review. In <i>The Thirty-ninth Annual Conference on Neural Information Processing Systems</i> , 2026. URL https://openreview.net/forum?id=z5KTxW5sJd .	557 558 559 560
Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. <i>Nature</i> , 651(8107):914–919, 2026.	561 562
Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. <i>arXiv preprint arXiv:2109.09774</i> , 2021.	563 564
Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. <i>arXiv preprint arXiv:2306.03262</i> , 2023.	565 566 567
Scott Page. <i>The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition</i> . Princeton University Press, 2008.	568 569

- Navita Goyal, Ivan Stelmakh, Nihar Shah, and Hal Daumé III. Causal effect of group diversity on redundancy and coverage in peer-reviewing. *arXiv preprint arXiv:2411.11437*, 2024.
- Walter J Rogan and Beth Gladen. Estimating prevalence from the results of a screening test. *American journal of epidemiology*, 107(1):71–76, 1978.
- Xingyao Wang, Simon Rosenberg, Juan Michelini, Calvin Smith, Hoang Tran, Engel Nyst, Rohit Malhotra, Xuhui Zhou, Valerie Chen, Robert Brennan, and Graham Neubig. The openhands software agent sdk: A composable and extensible foundation for production agents, 2026. URL <https://arxiv.org/abs/2511.03690>.
- J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990.
- Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.
- Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843, 2024.
- David Adam. The peer-review crisis: how to fix an overloaded system. *Nature*, 644(8075):24–27, 2025.
- Howard Bauchner and Frederick P Rivara. Use of artificial intelligence and the future of peer review. *Health Affairs Scholar*, 2(5):qxae058, 2024.
- Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.
- Miryam Naddaf. More than half of researchers now use ai for peer review—often against guidance. *Nature*, 649(8096):273–274, 2026.
- Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B Shah. Usefulness of llms as an author checklist assistant for scientific papers: Neurips’24 experiment. *arXiv preprint arXiv:2411.03417*, 2024.
- Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*, 2025.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, 2024.
- Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=CYmF38ysDa>.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL <https://openreview.net/forum?id=tLZZZiGpJX>.
- Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024.
- Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5877–5919, 2025.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In <i>The Thirteenth International Conference on Learning Representations</i> , 2025. URL https://openreview.net/forum?id=3GTtZFiajM .	621 622 623 624
Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? <i>Journal of Artificial Intelligence Research</i> , 75:171–212, 2022.	625 626
Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-agent review generation for scientific papers. <i>arXiv preprint arXiv:2401.04259</i> , 2024.	627 628
Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, et al. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 10164–10184, 2024.	629 630 631
Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, et al. Ai-driven review systems: evaluating llms in scalable and bias-aware academic reviews. <i>arXiv preprint arXiv:2408.10365</i> , 2024.	632 633 634
Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, et al. Reviewrl: Towards automated scientific review with rl. In <i>Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing</i> , pages 16942–16954, 2025.	635 636 637
Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 1208–1226, 2024.	638 639 640
Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. A large-scale randomized study of large language model feedback in peer review. <i>Nature Machine Intelligence</i> , pages 1–11, 2026.	641 642 643
Nicolas Bougie and Narimawa Watanabe. Generative reviewer agents: Scalable simulacra of peer review. In <i>Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track</i> , pages 98–116, 2025.	644 645 646
Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. Revieweval: An evaluation framework for ai-generated reviews. <i>arXiv preprint arXiv:2502.11736</i> , 2025.	647 648
Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. In <i>Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing</i> , pages 15662–15693, 2025.	649 650 651 652
Sarina Xi, Vishisht Rao, Justin Payan, and Nihar B Shah. Flaws: A benchmark for error identification and localization in scientific papers. <i>arXiv preprint arXiv:2511.21843</i> , 2025.	653 654
Hyun Ryu, Doohyuk Jang, Hyemin S Lee, Joonhyun Jeong, Gyeongman Kim, Donghyeon Cho, Gyouk Chu, Minyeong Hwang, Hyeongwon Jang, Changhun Kim, et al. Reviewscore: Misinformed peer review detection with large language models. <i>arXiv preprint arXiv:2509.21679</i> , 2025.	655 656 657
Xingyao Wang, Simon Rosenberg, Juan Michelini, Calvin Smith, Hoang Tran, Engel Nyst, Rohit Malhotra, Xuhui Zhou, Valerie Chen, Robert Brennan, et al. The openhands software agent sdk: A composable and extensible foundation for production agents. <i>arXiv preprint arXiv:2511.03690</i> , 2025.	658 659 660

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [CMUPaperReviewerappendix.pdf](#)