Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Frontier language models recover scientific quality but predictably diverge from human reviewers

**Vaibhava Lakshmi Ravideshik**
University of Southern California
**Zhisheng Tang**
University of Southern California
**Mayank Kejriwal**
kejriwal@isi.edu

University of Southern California

**Research Article**

**Keywords:** research paper evaluation, LLM peer review, publication venues, human-AI agreement, research quality metrics

**Posted Date:** November 12th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11132861/v1

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

**Additional Declarations:** No competing interests reported.

# Frontier language models recover scientific quality but predictably diverge from human reviewers

Vaibhava Lakshmi Ravideshik[1], Zhisheng Tang[1], Mayank Kejriwal[1*]

[1]University of Southern California, Los Angeles, CA, USA.

*Corresponding author(s). E-mail(s): kejriwal@isi.edu;
Contributing authors: vlds@umich.edu; zhisheng@usc.edu;

**Abstract**

Frontier language models are now deployed for peer review at scale, yet whether they recover established scientific quality hierarchies and how they diverge from human reviewers remain unclear. We evaluate 1,066 machine learning papers from 2025 main artificial intelligence conferences (NeurIPS, ICML, ICLR), matched workshops, and citation-stratified journals using three frontier models (GPT 5.4, Gemini 3.1 Pro, Claude Opus 4.6) in consensus on a 4-point reviewing rubric. Consensus scores recover every expected quality signal: main-conference papers outscore workshops by 0.40 units ($\boldsymbol{p = 3.93 \times 10^{-55}}$), presentation rankings emerge without access to acceptance metadata, and highly cited journal papers substantially outperform lowly cited counterparts. However, all three models converge on suppressing originality by 14–17 percentage points, and score variation across research areas reflects cross-model disagreement patterns. In a nested analysis of 162 ICLR papers, model-human alignment differs sharply between paper strengths (high agreement) versus weaknesses (low agreement). While frontier models agree with one another more than human reviewers agree with each other, strength-weakness alignment is only weakly coupled ($\boldsymbol{r \approx 0.17}$–$\boldsymbol{0.19}$), indicating this asymmetry is structural to review itself rather than an artifact of AI deployment. These findings establish that multi-model consensus recovers meaningful quality signals, but deployments require explicit accounting for originality underweighting and persistent gaps in weakness detection.

**Keywords:** research paper evaluation, LLM peer review, publication venues, human-AI agreement, research quality metrics

# 1 Introduction

Scientific peer review is under acute structural pressure, especially in artificial intelligence (AI) and machine learning. For its 2025 edition, the International Conference on Representation Learning (ICLR) — a premier machine learning conference with a current h5-index of 362 — saw a 61% submission increase between 2024 and 2025 [11]. More than half of 1,600 surveyed academics now report using AI tools and frontier LLMs during peer review, even against explicit journal guidance [12]. By 2024, Liang et al. had already documented that 6.5–16.9% of major conference review text was substantially LLM-modified [7, 8]. More recently, over 21% of ICLR 2025 reviews showed signs of AI generation [11].

In response, major venues like ICLR and AAAI (h5-index = 232) launched parallel pilots in 2026 involving AI reviewing [15]. But despite its clear importance as a modern phenomenon to both meta-science and research ethics, many open questions remain surrounding AI-generated reviews. Emerging approaches to systematic paper evaluation employ large language models (LLMs) to augment or automate the peer review process. Recent work has established foundational frameworks for automated scholarly paper review [10], enabling systematic quality assessment at scale. Previous studies have documented the potential of LLMs to provide useful feedback on research papers [9], though questions remain about their reliability as reviewers [19]. Multi-agent LLM frameworks have shown promise in modeling peer review dynamics [5], and recent standardized approaches decompose the evaluation process into distinct stages [18]. The broader shift toward LLM-assisted evaluation has been comprehensively surveyed [21], and controlled trials now show that LLM-generated feedback can measurably improve aspects of review quality at scale [16].

The literature leaves several critical gaps. First, it remains unclear whether frontier LLMs trained on diverse internet data can recover the *quality hierarchies* embedded in peer review systems—that is, whether they converge on similar assessments as human experts and reproduce known quality distinctions across venues, presentation formats, and citation impact. Second, prior work has not systematically characterized how different models weight different evaluation dimensions (e.g., *originality* versus *clarity*); do all frontier models exhibit the same biases, or do systematic dimensional preferences emerge across models? Third, content-level analysis of model-human agreement on specific strengths and weaknesses is sparse: agreement on numerical scores alone does not necessarily translate to agreement on the justifications underlying those scores. Finally, it is unknown whether observed divergence between machines and humans reflects AI-specific limitations or more fundamental structural properties of peer review itself.

To address these gaps, we structure this work around four research questions:
**RQ1: Do frontier LLMs empirically recover established quality hierarchies?** We ask whether consensus scores from multiple frontier models reproduce the quality distinctions encoded in peer review systems across venue type, presentation format, and citation impact, and how cross-model agreement varies across research domains.
**RQ2: Do frontier models systematically bias their evaluations along specific dimensions?** We ask whether independent LLMs converge on over- or under-weighting particular criteria relative to an unbiased baseline.

**RQ3: How do machine and human reviews align at the content level?** We ask whether consensus LLM reviews and human reviews show comparable patterns of agreement and disagreement when assessed on specific enumerated strengths and weaknesses rather than just agreement on numerical scores.
**RQ4: Is machine-human divergence AI-specific or structural to review?** We ask whether observed asymmetries (e.g., in strength versus weakness detection) are unique to AI systems or reflect intrinsic properties of peer review shared by both human and machine reviewers.

To answer these questions, we systematically gather 1,066 machine learning papers from three top-tier 2025 main conferences (NeurIPS, ICML, ICLR), topically matched workshops, and citation-stratified journal publications. Using three independent frontier models (GPT 5.4, Gemini 3.1 Pro, Claude Opus 4.6) without access to acceptance metadata or venue information, we deploy a consensus-based reviewing process on a standardized 4-point rubric. We assess whether consensus scores correlate with established quality signals and characterize how research area influences both score level and inter-model agreement.

A key limitation of venue-level benchmarking alone is that score validity does not guarantee content-level alignment with human judgment. To address this gap, we embed a nested analysis within the corpus: a content-level comparison between consensus reviews and human reviews for 162 accepted ICLR 2025 main-conference papers drawn from the larger dataset, for which anonymized human review text is available. This nested design allows us to ask not only whether LLM reviewers recover large-scale quality structure, but also where they converge with and diverge from human reviewers at the specific level of strengths and weaknesses. Together, these analyses provide systematic benchmarks for AI-reviewer behavior across venue types, research areas, citation impact levels, and review-content dimensions, with direct implications for AI-assisted peer reviewing in empirical machine learning research.

# 2 Methods

## 2.1 Dataset construction and composition

We document this dataset following the *Datasheets for Datasets* framework [3], which structures dataset documentation across motivation, composition, collection, preprocessing, uses, distribution, and ethical considerations.

The overall data-and-review pipeline is schematized in Figure 1. We compiled a comprehensive dataset of 1,066 research papers stratified across publication venue and citation impact (Table 1). The dataset comprises two primary components: (1) 866 papers from 2025 machine learning conferences and associated workshops, and (2) 200 journal articles published in 2021 with stratified citation profiles. Conference and workshop papers were sourced from three premier machine learning venues in 2025. NeurIPS contributed 146 main conference and workshop papers each (across two workshops). Similarly, ICML and ICLR respectively contributed 125 and 162 main conference and workshop papers each (also across two workshops). In total, this yielded a set of 433 main conference and workshop papers each across six distinct workshops

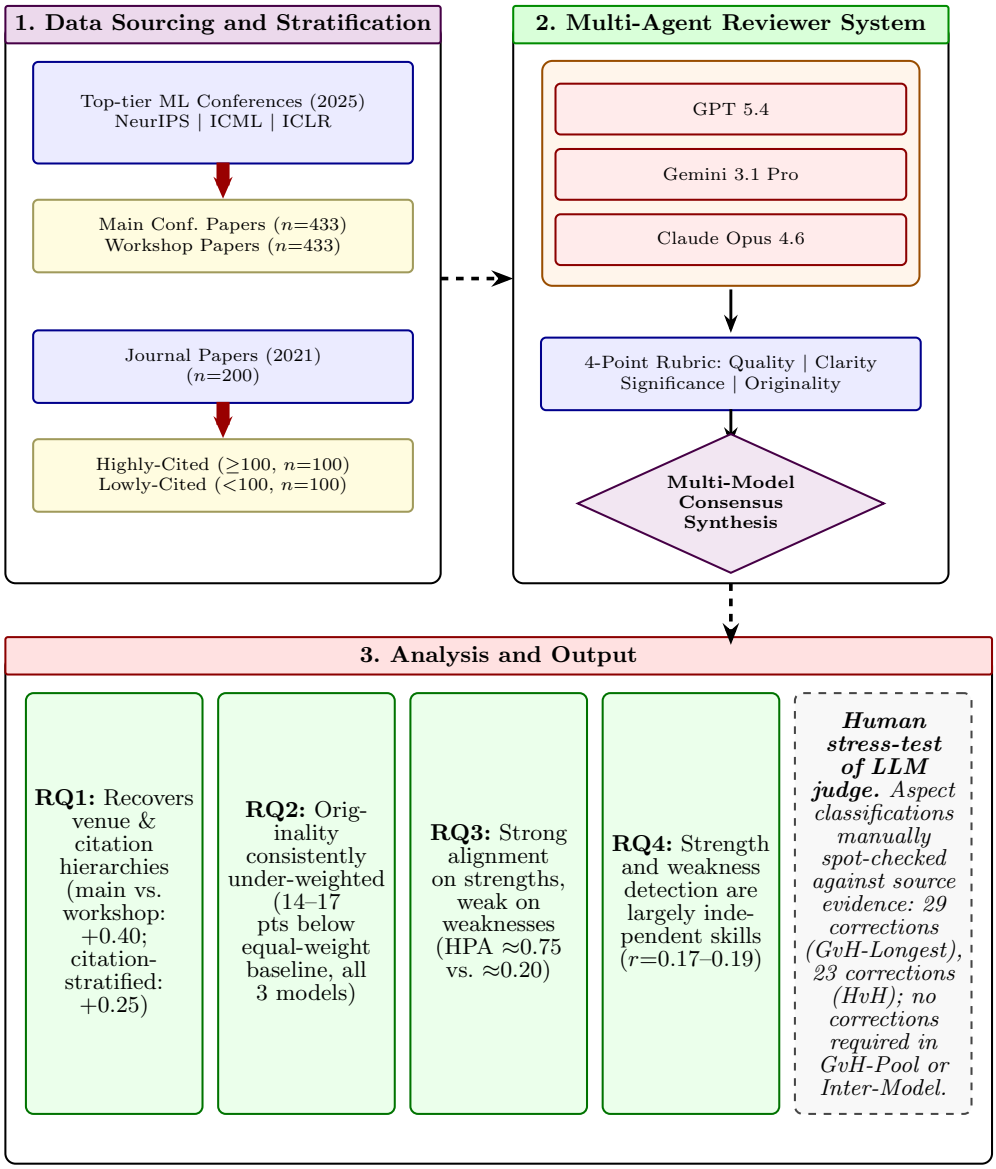



**Fig. 1** Architecture of the multi-model consensus reviewer system. The workflow comprises three stages: (1) Data Sourcing and Stratification, where papers are collected from top-tier ML conferences (NeurIPS, ICML, ICLR) in 2025 and journal publications from 2021, stratified by citation impact; (2) Multi-Agent Reviewer System, employing three independent frontier language models (GPT 5.4, Gemini 3.1 Pro, Claude Opus 4.6) that conduct parallel assessments using a 4-point rubric across four dimensions (Quality, Clarity, Significance, Originality), followed by multi-model consensus synthesis; and (3) Analysis and Output, structured around the paper's four research questions, with each stage producing dimensional scores, written analyses, confidence metrics, and ethical assessments. All aspect-level classifications underlying RQ3–RQ4 were manually spot-checked against source evidence, yielding 29 corrections in GvH-Longest and 23 in HvH.

**Table 1 Dataset Composition and Selection Strategy.** Overview of paper collection across three venue types (main conferences, workshops, journals) with 1,066 papers total. Conference and workshop papers were sourced from 2025 venues; journal papers were from 2021 with stratified citation impact.

| **Venue Type** | **Source** | **Papers** |
|---|---|---|
| Main Conferences | NeurIPS 2025 | 146 |
| | ICML 2025 | 125 |
| | ICLR 2025 | 162 |
| Workshops | 6 workshops (3 venues) | 433 |
| Journal Papers | Highly cited ($\geq$100 citations, 2021) | 100 |
| | Lowly cited ($<$100 citations, 2021) | 100 |
| **Total Dataset** | | **1,066** |

**Table 2 Workshop dataset composition.** Six workshops from three conferences totaling 433 papers. All workshops cluster thematically around geometric representations, generative models, and representational learning.

| **Conference** | **Workshop** | $n$ |
|---|---|---|
| NeurIPS 2025 | NeurReps (4th annual) | 121 |
| | NEGEL (1st annual) | 25 |
| ICML 2025 | HiLD (3rd annual) | 84 |
| | SIM (2nd annual) | 41 |
| ICLR 2025 | DeLTa (1st annual) | 125 |
| | Re-Align (2nd annual) | 37 |
| **Total** | | **433** |

(Table 2), enabling direct paired comparison of the same research topics across venue types while controlling for temporal and disciplinary factors.

The six workshops cluster thematically around three primary research areas: (1) geometric and topological deep learning, encompassing symmetry-aware representations, non-Euclidean spaces, and equivariance principles; (2) generative model theory and practice, including deep generative models and their theoretical underpinnings; and (3) representational learning and alignment across biological and artificial neural systems. As a result, our workshop sample may not generalize to workshops in other machine learning domains such as natural language processing, computer vision, or reinforcement learning applications.

Journal papers (200 total) were identified through systematic searches on Google Scholar using a curated vocabulary of 50 keywords extracted from the collected conference and workshop papers. Papers were restricted to those published in 2021, creating a deliberately designed temporal separation from the 2025 conference papers. Papers published in 2021 have accumulated approximately five years of citation history as of 2026. Books and review papers were excluded to focus on original research contributions. Within the journal sample, papers were stratified by citation impact using a 100-citation threshold: for each of the 50 keywords, we identified the first two published articles meeting the highly cited threshold ($\geq$ 100 citations) and separately the first two articles meeting the lowly cited threshold ($<$ 100 citations), producing exactly 100 papers per citation category.

Each paper record in the dataset includes: (1) individual dimensional scores (*Quality, Clarity, Significance, Originality* on a 4-point rubric) from each of three frontier language models; (2) multi-model consensus and composite scores; (3) complete text of written review analyses, identified questions, and limitations from each model; (4) confidence ratings from each model; (5) ethical concern flags and descriptions; (6) venue metadata; and (7) citation impact category for journal papers. Five papers (four main conference, one lowly cited journal) have incomplete individual model reviews due to processing failures but proceeded to consensus synthesis using available outputs.

The content-level human-alignment analysis needed for RQs 3 and 4 uses all 162 ICLR 2025 main-conference papers, which form a strict subset of the 1,066-paper corpus. This nested design lets us reuse the same reviewed manuscripts and consensus outputs while adding human-review comparisons only where OpenReview review text is available in a standardized form.

## 2.2 Collection process and multi-model reviewer assessment system

All 1,066 papers were submitted to the multi-model consensus reviewer as complete PDF manuscripts [10]. The system employs three independent frontier language models — GPT 5.4 (OpenAI), Gemini 3.1 Pro (Google), and Claude Opus 4.6 (Anthropic) — conducting parallel assessments without access to each other's evaluations. Each paper receives independent evaluation across four quality dimensions on a 4-point rubric (1=Poor, 2=Fair, 3=Good, 4=Excellent): (1) **Quality**, assessing technical soundness and methodological rigor; (2) **Clarity**, evaluating organization and writing quality; (3) **Significance**, assessing potential impact; and (4) **Originality**, examining novelty and differentiation from prior work. Beyond dimensional scoring, the system generates written analyses, targeted questions and limitations, confidence ratings, and ethical concern evaluations [18]. Following independent review, assessments are synthesized through multi-model consensus procedures [5]. All papers were processed between May and June 2026.

For each main conference paper, we assigned a presentation type label (Oral, Spotlight, or Poster) by querying the OpenReview API using the paper's forum ID. The `venue` field returned by the API contains a string encoding both the conference name and the presentation format (e.g., "ICLR 2025 Poster", "NeurIPS 2025 spotlight", "ICLR 2025 Oral"); we parsed this string to extract the presentation type label. Of the

433 main conference papers, 20 were labeled Oral, 40 Spotlight, and 318 Poster, with 55 papers receiving an "Unknown" label in cases where the OpenReview venue string contained no presentation type suffix despite multiple API endpoint attempts. Presentation type labels were not assigned to workshop papers, as workshop OpenReview metadata does not consistently encode this distinction. Journal papers were assigned the label "Journal Article" as presentation type is not applicable to that stratum.

## 2.3 Preprocessing and metadata enrichment

Papers were submitted as original PDFs without modification. Upon completion of review, we extracted and stored: (1) individual dimensional scores from each model; (2) consensus and composite scores; (3) complete written analyses; (4) confidence ratings; and (5) ethical concern flags. All outputs were imported into a structured database indexed by paper identifier, venue type, publication year, and citation impact category. No score normalization was applied prior to analysis; raw 4-point rubric scores are preserved in the released dataset.

## 2.4 Statistical analysis for venue-level benchmarking

To investigate RQ1 and RQ2, we conducted the following analyses. To quantify multi-model agreement, we computed cross-model standard deviation across GPT 5.4, Gemini 3.1 Pro, and Claude Opus 4.6 composite scores for each paper, where composite score is defined as the mean of the four dimensional scores (Quality, Clarity, Significance, Originality). Our primary quality metric is the consensus composite score, computed as the mean of the four consensus-synthesized dimensional scores. Venue-level comparisons were conducted using Welch two-sided $t$-tests with bootstrap confidence intervals (10,000 resamples). For citation-stratified journal pairs, paired two-sided $t$-tests were applied to the 100 matched highly cited/lowly cited pairs sharing the same keyword. All tests were conducted at $\alpha = 0.05$.

For dimensional attention analysis, each model's composite score was regressed on that model's four dimensional scores. Positive standardized coefficients were normalized to sum to 100%, yielding a regression-implied attention share per dimension. Deviations from the 25% equal-weight baseline quantify each model's systematic over- or under-weighting of each criterion. Residual burden was assessed as RMSE between observed and fitted scores. Area-level disagreement is the mean between-model standard deviation across papers within each annotated primary area; areas required at least three annotated papers.

## 2.5 Nested content-level comparison with human reviews

To investigate RQ3 and RQ4, we analyzed all 162 ICLR 2025 main-conference papers from the larger corpus, with official reviews retrieved via the OpenReview API (mean 4.0 reviewers per paper, range 3–6). Consensus reviews were generated by the same three frontier LLMs (GPT 5.4, Gemini 3.1 Pro, Claude Opus 4.6) operating independently without access to each other's outputs or the human reviews, before synthesis into a unified consensus output. An independent LLM judge (GPT-4o, temperature 0)

extracted 5–8 major strengths and weaknesses per review source and labeled each aspect yes/no for every other source.

Throughout, the *focal source* is the reviewer or system whose coverage is being evaluated: in GvH configurations the consensus reviewer is focal; in Inter-Model each model serves as focal in turn against the other two; in HvH each human reviewer is focal against the remaining reviewers on the same paper.

Four configurations were evaluated: **GvH-Longest** (consensus vs. the single longest human review per paper, by character count), **GvH-Pool** (consensus vs. all human reviewers simultaneously in a single judge call), **Inter-Model** (the three individual model reviews against one another prior to consensus synthesis), and **HvH** (human reviewers against each other without reference to AI outputs, establishing a human inter-reviewer baseline). Aspects were classified via pooled-majority rule into Common (focal source and at least two others agree), focal-unique (only the focal source raises it), and reference-unique (raised by reference sources but not the focal source); the threshold of two was chosen to require genuine cross-source consensus rather than coincidental pairwise agreement. From these we computed three metrics:

$$\text{HPA} = \frac{\text{Common}}{\text{Common} + \text{reference-unique}} \tag{1}$$

$$\text{CI} = \frac{\text{focal-unique}}{\text{Total}} \tag{2}$$

and, for the symmetric HvH and Inter-Model configurations where no single focal source is designated:

$$\text{Agreement Rate} = \frac{\text{Common}}{\text{Total}} \tag{3}$$

Reviewer positions R5 and R6 were excluded from reviewer-level analyses due to fewer than 50 valid aspect-level observations each. Reviewer-level HPA-Strengths and HPA-Weaknesses were computed per reviewer–paper pair for the HvH and GvH-Pool configurations respectively, after excluding pairs with undefined HPA denominators (arising when a reviewer raised no aspects on a given dimension). All classifications were spot-checked against source evidence by the first author; corrections were applied where the assigned type was inconsistent with the underlying yes/no pattern (29 corrections in GvH-Longest, 23 in HvH; no corrections were required in GvH-Pool or Inter-Model). Strength-versus-weakness differences used one-sided Mann–Whitney $U$ tests; reviewer-level coupling used Pearson correlation ($\alpha = 0.05$).

## 2.6 Dataset Use, ethics, and maintenance

The dataset is suitable for research on automated peer review, venue-specific quality benchmarking, citation impact analysis, multi-agent LLM evaluation, and analysis of reviewing style differences between LLM and human reviewers. The dataset should not be used to make definitive accept/reject decisions on papers, to rank or compare individual authors, or to draw conclusions about workshop domains not represented in the sample. All papers analyzed are publicly available research. No human subjects data or personally identifiable information was collected or used. Author identities were not

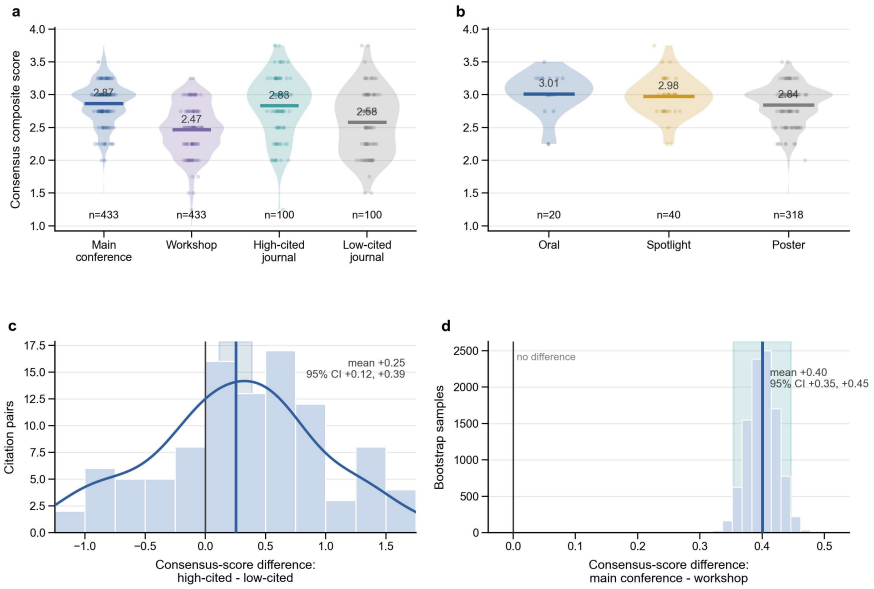



**Fig. 2 Consensus-Score Distributions Reveal Systematic Quality Differentiation Across Paper Types and Presentation Formats. a,** Cohort-level consensus-score distributions show that main-conference papers and highly cited journal papers occupy the higher score range, whereas workshop and lowly cited journal papers are lower on average. Mean consensus scores were 2.87 for main-conference papers, 2.47 for workshop papers, 2.83 for highly cited journal papers and 2.58 for lowly cited journal papers. **b,** Main-conference presentation-label distributions compare only mutually comparable main-conference records with explicit presentation labels. Oral and spotlight papers had higher mean consensus scores than posters. Mean consensus scores were 3.01 for oral papers (n=20), 2.98 for spotlight papers (n=40) and 2.84 for posters (n=318). The remaining 55 main-conference records had unknown presentation labels and were excluded from this panel. **c,** The paired citation contrast shows that highly cited journal papers scored higher than their topic-matched lowly cited counterparts by 0.25 consensus-score units on average (bootstrap 95% CI, 0.12 to 0.39; paired two-sided $p=4.83 \times 10^{-4}$; n=100 pairs). **d,** Because conference and workshop papers are not organized as matched pairs, the conference-workshop analogue is shown as an unpaired bootstrap distribution of the mean consensus-score difference. Main-conference papers scored higher than workshop papers by 0.40 consensus-score units on average (bootstrap 95% CI, 0.35 to 0.45; Welch two-sided $p=3.93 \times 10^{-55}$; n=433 per group). Together, these four panels show that the AI-reviewer consensus score recovers familiar external and internal selectivity structure across paper types.

used as variables in any analysis. Consensus scores reflect automated assessments by frontier language models and may reproduce or amplify biases present in those models' training data; scores should not be used as a substitute for human peer review in editorial decisions [1, 13]. The complete dataset is publicly available at https://github.com/vaibhavalakshmiravideshik/AI-Reviewer-Dataset/tree/main; individual papers remain subject to their original copyright terms.

# 3 Results

## 3.1 RQ1: Consensus scores recover established quality hierarchies and area-level agreement patterns

Consensus scores reproduced every quality hierarchy the models had no access to at inference time (Fig. 2). Panel *a* shows score distributions across all four cohorts with clear separation. The conference-versus-workshop gap is statistically decisive: main-conference papers scored 0.40 units higher (mean 2.87 vs. 2.47; 95% CI 0.35–0.45; $p = 3.93 \times 10^{-55}$; $n = 433$ per group). The effect was consistent across venues: NeurIPS (2.91 vs. 2.51), ICML (2.84 vs. 2.43), and ICLR (2.86 vs. 2.47), confirming that no single conference drives the result. Panel *d* shows the bootstrap distribution of this difference, with the null reference far outside the observed range. Within the accepted pool, panel *b* shows scores ranking Oral above Spotlight above Poster ($3.01 > 2.98 > 2.84$), a monotone ordering recovered from content alone without acceptance metadata. Panel *c* provides a temporally independent validation: across 100 keyword-matched journal pairs, 63 of 100 pairs show a positive score difference (highly cited minus lowly cited), with the distribution shifted rightward (mean +0.25; 95% CI 0.12–0.39; $p = 4.83 \times 10^{-4}$), confirming that consensus scores track long-run citation impact.

Beyond venue and citation effects, research area also shapes both score level and inter-model agreement. Panel *c* of Fig. 3 plots mean consensus score against mean cross-model disagreement for each primary research area, revealing a partial orthogonality between quality and evaluability that aggregate scores obscure. Neuroscience and cognitive science occupies the high-score, low-disagreement quadrant: models rate these papers highly and agree on it. Learning theory scores equally high yet generates substantially greater cross-model disagreement, consistent with theoretical depth requiring domain expertise that diverges across architectures. Deep learning, the largest and most methodologically heterogeneous area, shows above-median disagreement at only a near-median score. Inter-model standard deviation carries reliability information the consensus score does not, and should be reported alongside it. Practically, this means that automated triage systems should flag high-disagreement papers for mandatory human review rather than routing them directly on the basis of a single consensus number.

## 3.2 RQ2: All three models exhibit convergent suppression of originality

Panel *F* of the dimensional-weighting analysis (Fig. 4) reveals striking convergence: all three models over-weight Quality (GPT: +13, Gemini: +16, Claude: +10 percentage points above the equal 25% baseline) and Significance (+8, +12, +8 points), and under-weight Clarity (−4, −7, −8 points). The largest deviation is Originality suppression: all three discount it by 14–17 points (GPT: −17, Gemini: −17, Claude: −14). Concretely, a paper that scores 1.0 on Originality but 4.0 on the other three dimensions would receive the same composite under the model-implied weighting as a paper scoring approximately 2.4 uniformly, masking a genuine quality difference. Originality is the criterion peer review guidelines universally identify as central to scientific

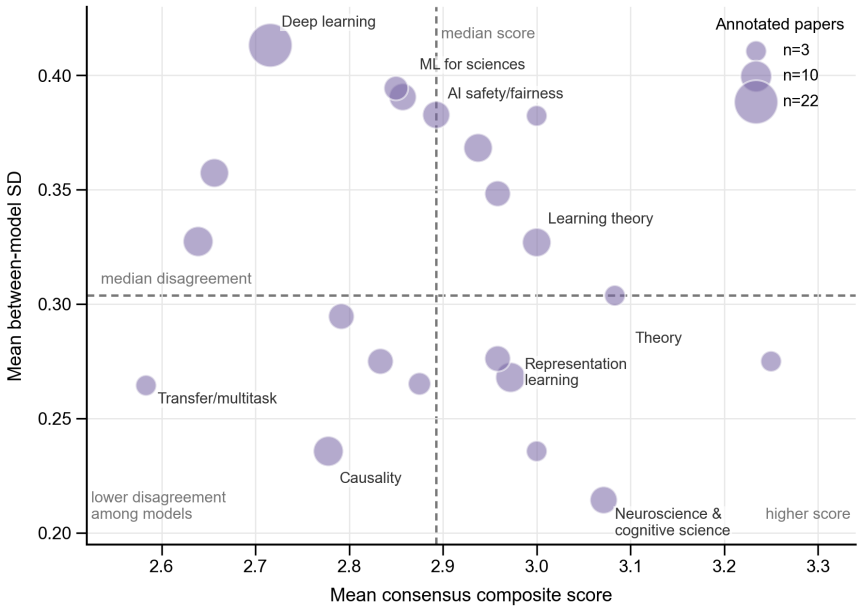



**Fig. 3 Primary Research Areas Show Distinct Patterns of Consensus Score and Model Agreement.** Primary-area annotations are from main-conference papers, and the plot includes primary-area groups with at least three annotated papers. Each point represents one annotated primary area. The x-axis is the mean consensus composite score across papers assigned to that area. The y-axis is the mean between-model standard deviation, calculated first within each paper as the standard deviation of the GPT, Gemini and Claude composite scores and then averaged across papers in the area; lower values therefore indicate stronger agreement among model reviewers, whereas higher values indicate greater model disagreement. Point size is proportional to the number of annotated papers in the area. Dashed lines mark the median area-level score and median area-level disagreement across the plotted areas. This separates average quality from model agreement: high-scoring areas can have either low disagreement, such as neuroscience and cognitive science, datasets and benchmarks, and representation learning, or higher disagreement, such as learning theory, reinforcement learning, optimization and interpretability. Broad heterogeneous areas such as deep learning show comparatively high model disagreement despite not having the highest mean consensus score. Thus, the figure suggests that primary areas differ not only in average model-reviewer score, but also in the stability of that evaluation across models; it should be interpreted as variation in the model-reviewer ensemble rather than as an objective ranking of research areas.

contribution. Its near-identical suppression across independently developed models points to a shared prior in scientific corpora, where technical soundness is discussed far more than novelty. This pattern is consistent with recent focus-level analyses showing that LLM reviewers systematically overweight technical validity while underweighting novelty relative to human reviewers [14, 17]. GPT showed greater internal scoring consistency than Gemini or Claude (RMSE 0.40 vs. 0.46 and 0.48), indicating that its composite scores diverge less from the dimensional average, but the directional bias in weighting is shared across all three.

To test whether these large-scale quality signals also correspond to the specific evaluative content that human reviewers discuss, we next turn to a deeper analysis of the nested ICLR subset.

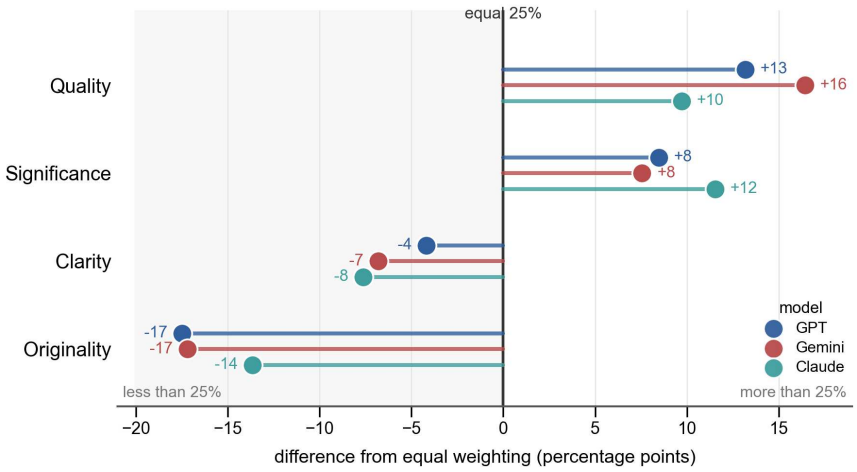



**Fig. 4 Model-Specific Attention Weights Reveal Systematic Differences in Dimensional Emphasis During Scoring.** For each scorer, the observed overall score was regressed on that scorer's quality, clarity, significance and originality scores: overall $= \beta_0 + \beta_1$quality $+ \beta_2$clarity $+ \beta_3$significance $+ \beta_4$originality. Standardized positive coefficients were normalized to sum to 100% within each model, giving a regression-implied attention share for each dimension. The main panel plots the direct difference between this attention share and the 25% equal-weight baseline. Positive values indicate more attention than equal weighting; negative values indicate less attention than equal weighting. Residual burden was lower for GPT than for Gemini and Claude: 20% of GPT-scored papers had $|\text{observed} - \text{fitted}| \geq 0.5$, compared with 31% for Gemini and 30% for Claude (RMSE = 0.40, 0.46 and 0.48, respectively).

## 3.3 RQ3: Machine-human alignment is asymmetric between strengths and weaknesses, and frontier LLMs agree with each other more than humans do

The consensus reviewer's alignment with pooled human reviewers (GvH-Pool) is markedly higher for strengths than weaknesses. Strength-flagged aspects cluster toward high HPA (corpus mean $\approx 0.75$, CI $\approx 0.26$): the consensus reviewer recovers most human-identified strengths while adding only modest unique content. Weakness-flagged aspects cluster toward low HPA (mean $\approx 0.20$, CI $\approx 0.33$): the consensus reviewer misses roughly four of five human-identified weaknesses while still generating weakness content no human corroborated. Dense stacks at HPA $= 0$ and 1 (Fig. 5*A*) show this asymmetry is largely binary per paper — the consensus reviewer typically recovers a paper's weaknesses almost entirely or misses them almost entirely. This asymmetry echoes recent large-scale audits reporting that LLM-generated reviews systematically inflate ratings for weaker submissions while more closely tracking human judgment on stronger ones [20], and that LLM reviewers frequently overlook weaknesses documented by human reviewers even at scale [22].

Beyond alignment with humans, frontier LLMs also agree with one another substantially more than human reviewers agree with each other. Human reviewers show limited mutual agreement, especially on weaknesses: HvH agreement is $\mu_S = 0.42$ for

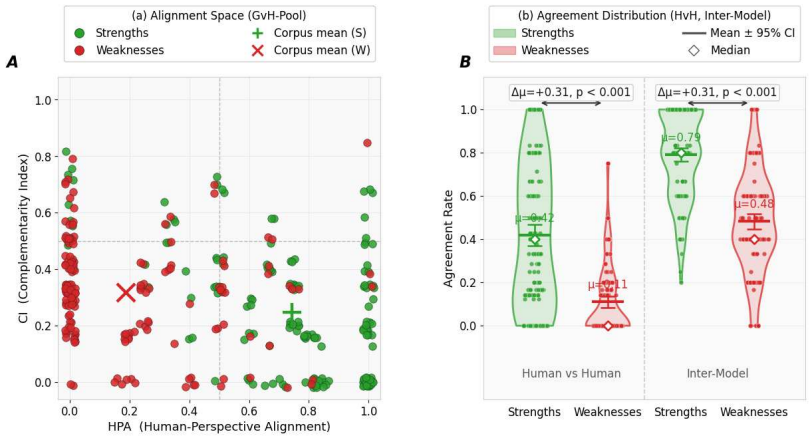



**Fig. 5** Alignment asymmetry between strengths and weaknesses, and agreement distributions across baselines. **(*A*)** GvH-Pool: consensus reviewer vs. pooled human reviewers across the full 162-paper corpus. Each point is one paper; green = strengths, red = weaknesses. $x$-axis: HPA = Common / (Common + Human-unique); $y$-axis: CI = AI-unique / Total. Corpus means (indicated by + for strengths, × for weaknesses) sit at approximately (HPA ≈ 0.75, CI ≈ 0.26) for strengths and (HPA ≈ 0.20, CI ≈ 0.33) for weaknesses. **(*B*)** Agreement distributions: human-vs.-human (HvH) vs. frontier-LLM inter-model agreement. HvH agreement is $\mu_S = 0.42$ vs. $\mu_W = 0.11$ ($\Delta\mu = +0.31$, $p < 0.001$); inter-model agreement is $\mu_S = 0.79$ vs. $\mu_W = 0.48$ (same $\Delta\mu = +0.31$, $p < 0.001$), exceeding HvH by 1.9× (strengths) and 4.4× (weaknesses).

strengths but $\mu_W = 0.11$ for weaknesses ($\Delta\mu = +0.31$, $p < 0.001$, Mann–Whitney $U$). The three frontier LLMs show the same directional asymmetry ($\mu_S = 0.79$ vs. $\mu_W = 0.48$, $\Delta\mu = +0.31$, $p < 0.001$), confirming weakness identification is intrinsically harder regardless of rater type (Fig. 5*B*). The two groups differ in absolute level, not shape: inter-model agreement exceeds HvH on both strengths (1.9×) and weaknesses (4.4×), showing frontier LLMs are far more internally consistent with each other than human reviewers are with one another.

## 3.4 RQ4: Strength-weakness decoupling is structural to peer review, not specific to AI

Two independent reviewer-level analyses converge on the same conclusion: a rater's ability to identify strengths is only weakly predictive of their ability to identify weaknesses. Considering human reviewers in isolation — each individual review compared against that paper's other human reviewers, with no AI or consensus reviewer involved — HPA-Strengths and HPA-Weaknesses are weakly correlated across 622 individual reviewer–paper pairs ($r = 0.173$, $p < 0.001$; corpus means $\mu_S \approx 0.56$, $\mu_W \approx 0.17$; Fig. 6*A*). The same weak coupling holds when comparing the consensus reviewer against each individual human reviewer across 588 reviewer–paper pairs ($r = 0.191$, $p < 0.001$; corpus means $\mu_S \approx 0.84$, $\mu_W \approx 0.35$; Fig. 6*B*). In both panels the point cloud sits almost entirely below the S = W diagonal, and the shallow OLS slope in each case shows that high strength-alignment carries little information about weakness-alignment. Because this pattern holds even in the purely inter-human comparison —

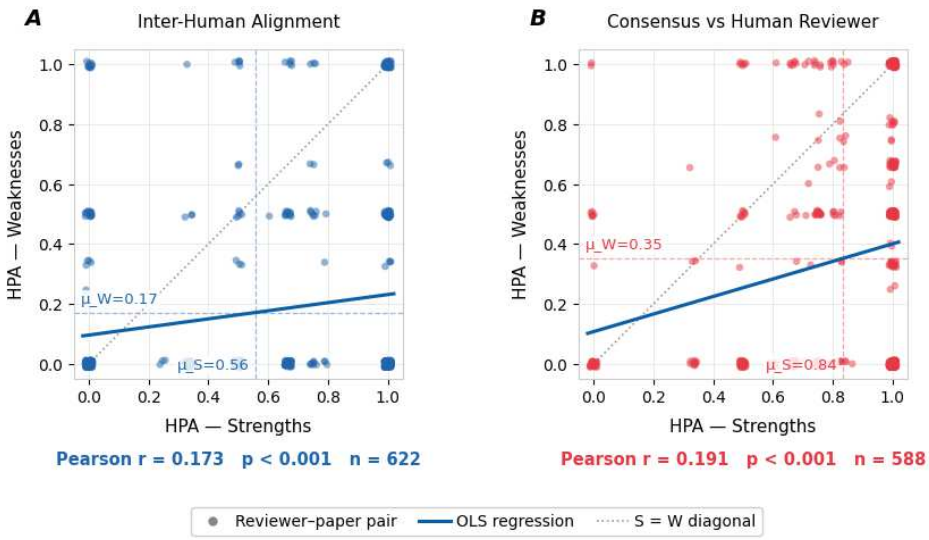



**Fig. 6** Strength and weakness alignment are weakly coupled at both the reviewer level and per-paper level. **(*A*)** Human reviewer–paper pairs ($n = 622$): HPA-Strengths vs. HPA-Weaknesses. Each point is one reviewer assessing one paper against that paper's other human reviewers; OLS regression line (blue) shows $r = 0.173$ ($p < 0.001$). **(*B*)** Consensus reviewer–paper pairs ($n = 588$): HPA-Strengths vs. HPA-Weaknesses, each point one reviewer per paper. OLS regression line shows $r = 0.191$ ($p < 0.001$). In both panels, the point cloud sits almost entirely below the S = W diagonal (dotted), and the shallow slope indicates that high strength-alignment carries little information about weakness-alignment.

where no AI system is involved at all — the strength–weakness decoupling appears to be a structural property of the review task itself, not an artifact of AI-human comparison.

# 4 Discussion

The results establish that multi-model LLM consensus recovers quality structure embedded in decades of selective publication without access to acceptance decisions, citation counts, or presentation metadata. This finding is consistent with recent large-scale benchmarking studies showing that frontier LLMs can match human reviewers on acceptance prediction [4] and correlate positively with citation impact [9], while extending those results to a multi-model, multi-venue, multi-stratum setting that prior work has not jointly addressed. Agreement across venue prestige, presentation format, and citation impact, at a scale that would require thousands of person-hours for human reviewers, provides convergent evidence that consensus scores encode genuine properties of scientific work. These findings extend a growing body of evidence that LLM-based review, when appropriately structured and validated, can approximate meaningful aspects of human evaluative judgment at a scale unattainable by manual review alone [16]. This positions multi-model consensus as a viable first-pass triage mechanism for high-volume review settings.

The research-area results reveal a structural limitation that aggregate scores obscure. Quality and evaluability are partially orthogonal: the same score carries less weight in a high-disagreement area such as learning theory than in a low-disagreement area such as neuroscience, making inter-model standard deviation a practical reporting necessity. Ignoring this dimension risks applying the same acceptance threshold uniformly across areas where the underlying measurement uncertainty differs substantially.

The Originality suppression finding carries the sharpest policy implications. The bias is convergent across three independently trained frontier models, the rubric explicitly assigned equal weight to all four dimensions, and no prompt-level explanation suffices; the suppression must originate in the models' training corpora, where empirical results are discussed more frequently than conceptual novelty [1, 13]. This is consistent with prior observations that LLMs tend to favor methodologically complete papers over conceptually novel ones [6], and with sociological accounts of how scientific corpora systematically under-represent novelty relative to validation [2]. Independent audits of LLM-assisted review pipelines have identified the same directional bias across different model families and evaluation frameworks [17]. Deployments of LLMs in peer review must correct for this bias or risk systematically disadvantaging the most creative work.

The nested human-review analysis clarifies both the promise and the limit of this apparent validity. On the one hand, the same system that recovers venue-level quality structure also aligns strongly with humans on strengths, suggesting that LLM reviewers reliably recognize many of the positive signals that selective publication systems reward. On the other hand, their much weaker alignment on weaknesses shows that successful score stratification does not imply equally strong critical diagnosis.

Strength identification behaves as a comparatively saturated task on which the consensus reviewer and its constituent LLMs approach or exceed the human inter-reviewer ceiling, while weakness identification remains the dominant source of both AI-human and human-human divergence — likely reflecting the domain judgment and critical scrutiny weakness detection requires, rather than an AI-specific deficiency. Prior work on human peer review has similarly documented that reviewer disagreement is substantially higher for weaknesses than strengths [2], and that critical diagnosis of limitations requires domain expertise that resists standardization [10]. Our finding that the strength–weakness gap is structurally identical in shape between human and AI reviewers, differing only in absolute magnitude, provides the first direct empirical evidence that this asymmetry is intrinsic to the review task rather than a failure mode of automated systems. This is consistent with, and extends, recent work showing that LLM reviewers diverge from human reviewers specifically in the content of identified weaknesses rather than in overall scoring calibration [20]. The weak strength–weakness coupling further indicates these are largely independent reviewing skills: a reviewer who reliably identifies strengths cannot be assumed to reliably identify weaknesses. AI-assisted review pipelines may therefore need distinct mechanisms — such as adversarial or role-specialized prompting — to close the weakness-detection gap rather than assuming strength-summarization gains transfer automatically.

While studies comparing AI reviewer scores across multiple models and venues with differential acceptance rates remain rare, our results are consistent with a population-based study that compared Stanford's Agentic Reviewer [4] scores to those of humans. In that study, the agentic reviewer was described as biased toward over-weighting technical validity over novelty [6]. Our results sharpen that concern by showing both a quantitative originality underweighting in large-scale scoring and a qualitative weakness-detection gap in content-level analysis, suggesting that current frontier systems are strongest at recognizing established merit signals and weakest at surfacing subtle novelty- and critique-related signals.

### Limitations

The workshop papers were drawn from venues topically aligned with the conference corpus, which may partially attenuate the observed venue gap. The content-level analysis is confined to a single venue and discipline, and it relies on an LLM judge whose biases may propagate despite manual audit. The highly cited/lowly cited journal stratification reflects citation accumulation over a specific five-year window that may not generalize to other fields or timescales. Broader replication across diverse disciplines, venues, and model generations remains an important next step.

## 5 Data availability

The complete dataset of 1,066 papers and consensus quality scores is publicly available at https://github.com/vaibhavalakshmiravideshik/AI-Reviewer-Dataset/tree/main. The content-level human-alignment derivatives used for the nested ICLR analysis will be packaged alongside the broader release upon acceptance.

## 6 Code availability

Statistical analysis code is available upon request.

## 7 Author contributions

Conceptualization, data collection, methodology, analysis, and writing were contributed by all authors.

## 8 Competing interests

The authors declare no competing interests.

## Appendix A Keyword vocabulary used for journal retrieval

The following 50 keywords represent the primary research areas, methodologies, and conceptual frameworks relevant to this study:

1. Geometric Deep Learning
2. Riemannian manifold
3. Equivariant neural networks
4. Persistent homology
5. Latent space geometry
6. Disentangled representations
7. Non-Euclidean geometry
8. Lie groups
9. Gauge equivariance
10. Product manifolds
11. High-dimensional learning dynamics
12. Neural tangent kernel
13. Random matrix theory
14. Causal representation learning
15. Spurious correlations
16. Invariant risk minimization
17. Out-of-distribution generalization
18. Interventional distributions
19. Structural causal models
20. Representational similarity analysis (RSA)
21. Centered kernel alignment (CKA)
22. Probing classifiers
23. Brain-score
24. Generalization bounds
25. Implicit regularization
26. Benign overfitting
27. Tensor decomposition
28. Probabilistic circuits
29. Double descent
30. Conjecture generation
31. Neural population geometry
32. Manifold untangling
33. Cross-modal alignment
34. Representational convergence
35. Universal representations
36. Progressive sharpening
37. Mean field theory neural networks
38. Flat minima generalization
39. Saddle point dynamics
40. Spiked covariance model
41. Marchenko-Pastur
42. Covariate shift
43. Latent variable disentanglement identifiability
44. Procrustes alignment
45. Geodesic distance representations

46. Shape metric neural networks
47. Functional similarity networks
48. Convergent learning
49. Margin theory neural networks
50. PAC-Bayes bounds

## References


[1] Biswas S, Dobaria D, Cohen HL (2023) ChatGPT and the future of journal reviews: A feasibility study. The Yale Journal of Biology and Medicine 96(4):415–420. https://doi.org/10.59249/TOMD6516

[2] Checco A, Bracciale L, Loreti P, et al (2021) AI-assisted peer review. Humanities and Social Sciences Communications 8(1):1–11. https://doi.org/10.1057/s41599-020-00703-8

[3] Gebru T, Morgenstern J, Vecchione B, et al (2021) Datasheets for datasets. Communications of the ACM 64(12):86–92. https://doi.org/10.1145/3458723

[4] Jiang Y, Ng A (2025) Stanford agentic reviewer. https://paperreview.ai

[5] Jin Y, Zhao Q, Wang Y, et al (2024) AgentReview: Exploring peer review dynamics with LLM agents. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, https://doi.org/10.48550/arXiv.2406.12708

[6] Kukoyi A (2026) Exploring stanford agentic reviewer: Equilibrium perspectives and the paradox of AI-driven peer review. SSRN 6112766, https://doi.org/10.2139/ssrn.6112766

[7] Liang W, Izzo Z, Zhang Y, et al (2024) Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In: Proceedings of the 41st International Conference on Machine Learning, pp 29575–29620

[8] Liang W, Zhang Y, Cao H, et al (2024) Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI 1(8):AIoa2400196. https://doi.org/10.1056/AIoa2400196

[9] Liang W, Zhang Y, Cao H, et al (2024) Can large language models provide useful feedback on research papers? a large-scale empirical analysis. NEJM AI 1(8):AIoa2400196. https://doi.org/10.1056/AIoa2400196

[10] Lin J, Song J, Zhou Z, et al (2023) Automated scholarly paper review: Concepts, technologies, and challenges. Information Fusion 98:101830. https://doi.org/10.1016/j.inffus.2023.101830

[11] Naddaf M (2025) AI is transforming peer review — and many scientists are worried. Nature 639(8056):852–854. https://doi.org/10.1038/d41586-025-00894-7

[12] Naddaf M (2026) More than half of researchers now use AI for peer review — often against guidance. Nature 649(8096):273–274. https://doi.org/10.1038/d41586-025-04066-5

[13] Robertson Z (2023) GPT-4 is slightly helpful for peer-review assistance: A pilot study. arXiv:2307.05492, https://doi.org/10.48550/arXiv.2307.05492

[14] Shin H, Tang J, Lee Y, et al (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 35618–35644

[15] Thakkar N, Yüksekgönül M, Silberg J, et al (2026) Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025. Nature Machine Intelligence https://doi.org/10.1038/s42256-026-01188-x, arXiv:2504.09737

[16] Thakkar N, Yuksekgonul M, Silberg J, et al (2026) A large-scale randomized study of large language model feedback in peer review. Nature Machine Intelligence 8(3):326–336

[17] Vasu SSM, Sheth I, Wang HP, et al (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

[18] Yu J, Ding Z, Tan J, et al (2024) Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp 10164–10184, https://doi.org/10.18653/v1/2024.findings-emnlp.595

[19] Zhou R, Chen L, Yu K (2024) Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, Italia, pp 9340–9351, URL https://aclanthology.org/2024.lrec-main.816/

[20] Zhu C, Xiong J, Ma R, et al (2025) When your reviewer is an LLM: Biases, divergence, and prompt injection risks in peer review. arXiv preprint arXiv:2509.09912

[21] Zhuang Z, Chen J, Xu H, et al (2025) Large language models for automated scholarly paper review: A survey. Information Fusion 124:103332

[22] Żurawicki K, Farganus J, Gaweł A, et al (2026) PRAIB: Peer review AI benchmark of behaviour of LLM-assisted reviewing. arXiv e-prints, arXiv-2605