Model-Specific Reviewer Styles in LLM Scientific Peer Review

Anonymous Authors
Anonymous Institution
anonymous@anonymous.edu

Abstract

Large language models are increasingly discussed as scientific reviewers, but they are still often treated as a single generic evaluator. We ask whether reviewer behavior is model-dependent when the protocol and human anchor are held fixed. We construct a benchmark centered on 250 OpenReview submissions with paper-level human score anchors, together with a 34-paper repeatability subset and a 30-paper prompt-robustness subset. We compare GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, and DeepSeek-V3.2. On the OpenReview benchmark, the models differ sharply in mean score bias, verdict usage, residual compression against human ratings, and rerun stability: Claude is closest to the human score anchor, GPT preserves more borderline decisions, Gemini is more permissive and less stable, and DeepSeek is highly permissive. Robustness analyses suggest that these patterns are not reducible to a single prompt wording or to a small number of title-recognizable manuscripts. Counterfactual reviewer-assignment analyses also show that model choice induces measurable decision-boundary shifts. LLM scientific reviewers should therefore be treated as distinct model-specific reviewer styles, not as interchangeable instances of a single "LLM reviewer."

1 Introduction

Large language models are increasingly used to critique papers, summarize reviews, and simulate referee behavior, but much of the discussion still treats "the LLM reviewer" as a single generic evaluator. Prior work has shown that strong models can generate plausible reviews or useful paper feedback (Liu and Shah, 2023; Gao et al., 2024; Liang et al., 2024), and newer systems optimize scientific reviewing with specialized prompting, search trees, or reinforcement learning (Idahl and Ahmadi, 2025; Chang et al., 2025; Zeng et al., 2025). Work on LLM-as-a-judge evaluation also shows that model judges can be biased, unstable, self-preferential, or prompt-sensitive rather than neutral scoring devices (Zheng et al., 2023; Ye et al., 2025; Zeng et al., 2024; Xie et al., 2025; Wang et al., 2024; Chen et al., 2025). Alignment work suggests a related concern: preference-trained assistants may over-weight agreement, encouragement, or user-congruent framing rather than purely truth-tracking responses (Sharma et al., 2023; Perez et al., 2022; Kaur, 2025; Kim and Khashabi, 2025; Jain et al., 2026). We start from a simple possibility: different frontier models may behave like different reviewers, with different strictness, boundary usage, and stability, even when asked to review the same paper under the same rubric.

We study this question in a shared human-anchored setting. The core benchmark contains 250 OpenReview submissions with paper-level mean human scores, plus a 34-paper rerun subset for repeatability and a 30-paper subset for prompt robustness. We also use Nature Communications and F1000 as supporting stress tests, but OpenReview is the only controlled human-anchored benchmark. The question is not just whether a model can produce plausible review text; it is whether model-specific reviewer styles can be measured against the same human anchor.

Across the OpenReview benchmark, the four models differ in average score, verdict usage, residual compression against human ratings, and rerun stability. Claude is closest to the human score anchor and the most stable; GPT is moderately more permissive but preserves many borderline decisions; Gemini is more permissive and much less stable; and DeepSeek is the most acceptance-oriented by a large margin. These patterns persist under prompt paraphrase, venue/year slice checks, and removal of a small set of title-recognizable manuscripts from a contamination-

risk audit. They also produce different counterfactual decision-boundary shifts when a model score is treated as an additional reviewer. The contribution is not a claim that one model is "the best reviewer," but a model-by-model measurement of scientific-review calibration, stability, and boundary behavior.

Related context. Our work connects scientific-review generation, LLM evaluation, and alignment-style behavior. Unlike automated-review systems, we do not propose a better reviewing agent; we benchmark how strong frontier models already differ under a shared human anchor. Unlike generic LLM-as-a-judge work, we focus on scientific peer review, where verdict boundaries, residual compression, and rerun stability matter as much as average score. This focus is also practical: recent field and randomized studies suggest that AI-assisted reviewing can shift paper scores or affect peer-review outcomes (Russo et al., 2025; Thakkar et al., 2026). Work on sycophancy and agreement-seeking offers one possible interpretation of model-level differences in criticism, encouragement, and context-following (Sharma et al., 2023; Perez et al., 2022; Kaur, 2025; Kim and Khashabi, 2025; Jain et al., 2026). We use that literature only as interpretation, not as proof of mechanism; the evidence in this paper is behavioral and benchmark-based. Here, reviewer style means a measurable pattern in score calibration, verdict-boundary usage, residual compression, and repeatability under a shared review protocol.

2 Experimental Setup

Models. We evaluate GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, and DeepSeek-V3.2. Exact API identifiers, access dates, decoding settings, and implementation details are provided in the supplementary material.

Prompt and outputs. All models receive the same structured reviewer prompt with the same score anchors and output schema. The prompt asks for a summary, strengths, weaknesses, soundness, presentation, contribution, rating, confidence, and a reject/borderline/accept verdict. Across providers, we normalize manuscript content, score anchors, and output schema as far as the available public interfaces allow.

Table 1: Core protocol summary. The study is paired at the paper level: each model reviews the same manuscripts under the same score anchors and output schema, and all model scores are compared to the same human-review anchor.

Component	Setting
Benchmark	250 OpenReview submissions with paper-level mean human scores
Models	GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, DeepSeek-V3.2
Model output	Structured review JSON with rating, confidence, and reject/borderline/accept verdict
Measures	Score bias, verdict usage, residual compression, rerun stability, and boundary shifts
Robustness	34-paper rerun subset, 30-paper prompt paraphrase subset, title/metadata audits, venue-slice checks
Artifacts	Exact prompts, API identifiers, decoding settings, and artifact pointers are in the appendix

Datasets. The OpenReview benchmark contains 250 submissions spanning ICLR 2024, NeurIPS 2024, ICLR 2025, and ICLR 2026. For each paper, we use the native numeric reviewer scores exposed by OpenReview and compute the paper-level mean human rating as the main anchor; we do not rescale these scores across venue-year slices. In the retained sample, the observed paper-level means happen to span 0 to 8, while the underlying individual reviewer scores occupy different observed subranges across venue-year slices; exact slice ranges are reported in the supplementary material. We also retain the full reviewer-score list for disagreement analysis. The rerun subset contains 34 papers evaluated three times per model, and the prompt-robustness subset contains 30 papers stratified across low-, mid-, and high-scored manuscripts. The Nature Communications and F1000 datasets are used only as supporting stress tests and are described in the supplementary material.

Leakage controls and scope. Table 1 gives the compact protocol. Because OpenReview manuscripts are public, we do not frame the benchmark as contamination-free; contamination risk is instead an auditable property of the setting. We reduce direct leakage in four ways: the prompt instructs models to ignore author, venue, and publication-status cues and to avoid external search; model inputs exclude reviewer comments, rebuttals, and decision metadata; standard-

ized files remove status cues where applicable; and we run venue/year-slice robustness checks, including exclusion of the full ICLR 2026 slice. We treat the result as a human-anchored behavioral benchmark of reviewer calibration. Supplementary title-only recognition probes, metadata-sufficiency controls, and high-risk exclusion analyses test whether the main ordering is driven by a small number of recognizable manuscripts.

Statistics. All primary OpenReview comparisons are paired at the paper level. We report mean differences with bootstrap confidence intervals and use paired Wilcoxon signed-rank tests for inter-model comparisons. We apply Benjamini–Hochberg FDR correction across the main pairwise model comparisons. For residual analyses, we report fitted slopes and compare them against a permutation null that preserves each model's marginal score distribution while randomizing paper identity.

3 OpenReview Results

3.1 Model-dependent reviewer behavior

Table 2 summarizes the 250-paper benchmark. Claude is closest to the human score anchor on average (5.088 vs. human 4.739, $\Delta = +0.349$, 95% bootstrap CI $[+0.217, +0.482]$), GPT-5.4 is moderately more permissive ($\Delta = +0.833$, CI $[+0.690, +0.973]$), Gemini is substantially more permissive ($\Delta = +1.445$, CI $[+1.225, +1.649]$), and DeepSeek is the most permissive by a large margin ($\Delta = +2.645$, CI $[+2.498, +2.807]$). Pairwise differences are also statistically strong: GPT differs from Claude ($p < 10^{-14}$, paired Wilcoxon), Claude differs from Gemini ($p < 10^{-21}$), and all comparisons involving DeepSeek are more extreme still. Figure 1 already hints that GPT's profile is not just a mean shift: its ratings remain visually concentrated around the middle of the scale, whereas Gemini and especially DeepSeek shift much more strongly toward high scores. The verdict distributions reveal distinct reviewer styles: GPT preserves many borderline decisions, Claude splits more evenly between borderline and accept, Gemini collapses toward accept while still issuing some hard rejects, and DeepSeek behaves as a highly accepting reviewer. This boundary behavior is not confined to the global average. When the papers are stratified by human score, GPT still assigns borderline

Table 2: OpenReview benchmark (250 submissions). Δ denotes model mean minus human mean. Verdict counts are reported as reject / borderline / accept. Larger positive Δ means stronger upward bias relative to the human anchor.

Model	n	Mean rating ↑	Δ vs. human	Verdicts
GPT-5.4	250	5.572	+0.833	9 / 147 / 94
Claude 4.6	250	5.088	+0.349	32 / 109 / 109
Gemini 3.1 Pro	250	6.184	+1.445	52 / 14 / 184
DeepSeek-V3.2	250	7.384	+2.645	2 / 20 / 228
Human	250	4.739	0.000	–

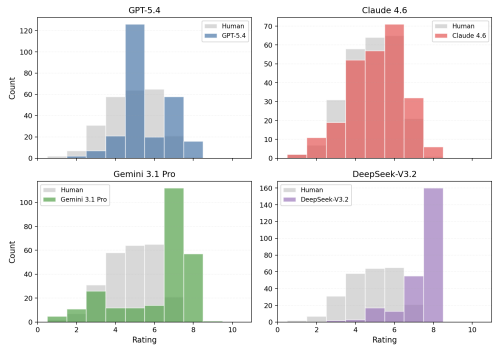

Figure 1: Human and model rating distributions on the 250-paper OpenReview benchmark. GPT-5.4 is visually concentrated around the middle of the scale, whereas Gemini and especially DeepSeek shift more strongly toward high scores.

to 78.3% of low-scored papers, 59.4% of mid-scored papers, and 32.1% of high-scored papers, whereas Gemini uses borderline on only 5.8%, 7.0%, and 1.9% of the same strata, and DeepSeek almost fully collapses to accept in the mid/high strata (96.9% and 100% accept, respectively). Figure 2 makes that bucketed pattern visible: GPT remains much more middle-concentrated than the other models even when the human anchor indicates clearly weak or clearly strong papers. Error-based summaries tell the same story. Against the paper-level human mean, Claude has the lowest MAE/RMSE ($0.898/1.135$), followed by GPT ($1.172/1.429$), Gemini ($1.922/2.246$), and DeepSeek ($2.655/2.932$).

3.2 Residual compression against human ratings

The residual plots in Figure 3 show that all four models exhibit a negative residual slope against human ratings. Lower-scored human papers are generally lifted upward by the models, while higher-scored papers are compressed downward or only modestly increased. This pattern is weak-

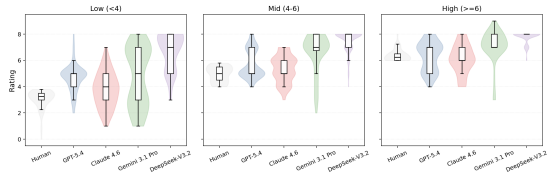

Figure 2: Model rating distributions within low-, mid-, and high-scored human-score buckets. GPT-5.4 remains more middle-concentrated across all three strata, whereas Gemini and DeepSeek move more decisively toward acceptance.

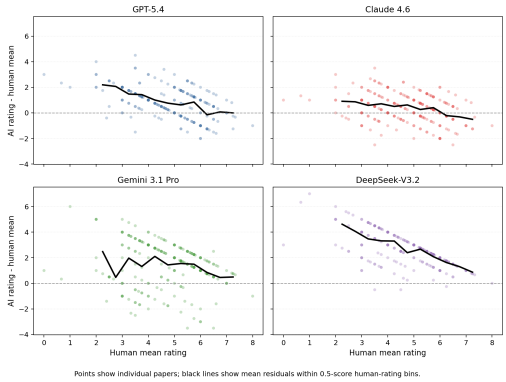

Figure 3: Residual analysis on the OpenReview benchmark. Points show individual papers and black lines show mean residuals within 0.5-score human-rating bins. All models lift weaker papers and compress stronger ones, but the strength of this pattern differs by model.

est for Claude (slope -0.252) and strongest for DeepSeek (slope -0.655), with GPT (-0.450) and Gemini (-0.277) in between. A trivial regression-to-the-mean explanation would predict slopes close to -1 if model scores were merely paper-independent draws with different marginal means. Under a permutation null that preserves each model's marginal score distribution, the 95% slope intervals are $[-1.112, -0.883]$ for GPT, $[-1.135, -0.863]$ for Claude, $[-1.167, -0.819]$ for Gemini, and $[-1.096, -0.905]$ for DeepSeek. All observed slopes are therefore substantially less negative than the null intervals, indicating structured reviewer behavior rather than a purely mechanical mean-mismatch effect.

3.3 Repeatability and human disagreement baseline

The 34-paper rerun subset reveals clear stability differences. Claude has the lowest mean rating standard deviation (0.100), followed by GPT (0.157) and DeepSeek (0.190), whereas Gemini is far less stable (0.804). Verdict consistency tells the same story: GPT, Claude, and DeepSeek all achieve an average majority-verdict ratio of at least 0.98, while Gemini drops to 0.843. This matters relative to ordinary human disagreement. Across the 250 papers, the mean within-paper human rating standard deviation is 1.174, the mean within-paper range is 2.432, and the mean absolute pairwise gap between human reviewers is 1.369. A stronger leave-one-reviewer-out baseline points in the same direction: comparing one human reviewer to the mean of the remaining reviewers yields MAE 1.199 and Pearson correlation 0.526. Against that baseline, Claude's MAE to the paper-level human mean is 0.898, GPT's is 1.172, while Gemini (1.922) and DeepSeek (2.655) are substantially worse.

3.4 Decision boundaries and reviewer-style fingerprints

The reviewer-style interpretation becomes sharper when we move from raw scores to decision boundaries. Using a simple human-anchor mapping of $< 4 \mapsto$ reject, 4–6 $\mapsto$ borderline, and $\geq 6 \mapsto$ accept, Claude has the strongest overall decision-level alignment (macro-F1 0.533, weighted $\kappa = 0.491$), whereas GPT has by far the highest borderline recall (0.594). Gemini and DeepSeek, by contrast, recover accept decisions much more readily than borderline or reject decisions: Gemini reaches accept recall 0.906 but borderline recall only 0.070, while DeepSeek reaches accept recall 1.000 but borderline and reject recall only 0.031 and 0.029. The same pattern appears in a simple counterfactual reviewer-assignment simulation. If one LLM score is added as an extra reviewer to the existing human scores, boundary-flip rates are 0.104 for Claude, 0.168 for GPT, 0.232 for Gemini, and 0.284 for DeepSeek. Among low-scored papers, the fraction pushed upward to borderline or accept is 0.044 for Claude, 0.060 for GPT, 0.092 for Gemini, and 0.152 for DeepSeek. Finally, the models differ in how often they stay within the observed human-reviewer score envelope: Claude falls within the human score range on 78.0% of papers, GPT on 73.6%, Gemini on 43.2%, and DeepSeek on 30.8%. Table 3 compresses the most decision-relevant dimensions into a compact reviewer-style summary; the full fingerprint appears in the supplementary material.

Table 3: Compact reviewer-style summary on the 250-paper OpenReview benchmark. Larger positive bias means stronger upward scoring relative to the human anchor. Lower low-score accept rate and lower counterfactual boundary-flip rate indicate a more conservative decision boundary, whereas higher borderline rate and higher in-range rate indicate stronger retention of intermediate judgments and closer positioning within the observed human-reviewer envelope.

Model	Bias Δ	Borderline rate ↑	Low-score accept ↓	Flip rate ↓	In-range rate ↑
GPT-5.4	+0.833	0.588	0.087	0.168	0.736
Claude 4.6	+0.349	0.436	0.101	0.104	0.780
Gemini 3.1 Pro	+1.445	0.056	0.478	0.232	0.432
DeepSeek-V3.2	+2.645	0.080	0.739	0.284	0.308

4 Robustness and Scope

Prompt robustness. To test whether the reviewer-style findings depend on one prompt wording, we ran a prompt-robustness check on a 30-paper stratified subset while keeping the task, score anchors, and output schema fixed and changing only the surface wording of the instructions. GPT and Claude remained highly stable: GPT's mean score shift was $+0.033$, its mean absolute error relative to human ratings changed by only -0.021, and its verdict-agreement rate was 0.967; Claude's corresponding numbers were -0.033, $+0.078$, and 0.933. DeepSeek showed a larger but still largely verdict-preserving shift, whereas Gemini showed larger movement that is difficult to separate from its broader rerun instability. Figure 4 summarizes the same pattern: GPT and Claude are close to wording-invariant under this mild paraphrase, whereas Gemini and DeepSeek are more sensitive. These results make a single-prompt explanation unlikely, while also showing that some models are more prompt-sensitive than others.

Contamination-risk audit. The audit is designed to bound how much public-manuscript recognizability could explain the results, not to prove absence of contamination. In a full-benchmark title-only audit using Claude as a strong recognizability probe, only $7/250$ papers were marked recognizable. On a four-model 50-paper title-only subset, the recognized counts are similarly low (GPT 1, Claude 4, Gemini 1, DeepSeek 0), so the recognizability signal is not a Claude-only artifact. On the same 50-paper subset, a metadata-sufficiency control using title plus abstract does not reproduce the full-manuscript behavior, especially for DeepSeek, whose mean score drops from 7.70 under full-manuscript evaluation to 5.62

Figure 4: Prompt robustness on the 30-paper OpenReview subset. Left: change in mean absolute error relative to the human mean under a paraphrased prompt. Right: verdict agreement across prompt variants.

Table 4: Compact robustness summary on the 50-paper contamination-audit subset. Metadata-only MAE diff measures the mean absolute score difference between title+abstract-only and full-manuscript evaluation (smaller is better). Text-only Δ is the mean model–human score gap under a standardized extracted-text pipeline. Residual slopes compare the full 50-paper subset with the 46-paper high-risk-exclusion subset; values closer to zero indicate weaker compression.

Model	Meta-only MAE diff ↓	Meta verdict agr. ↑	Text-only Δ	Residual slope (50→46, toward 0)
GPT-5.4	0.76	0.68	+0.731	$-0.539 \rightarrow -0.512$
Claude 4.6	0.56	0.78	+0.471	$-0.551 \rightarrow -0.548$
Gemini 3.1 Pro	1.36	0.70	+1.431	$-0.534 \rightarrow -0.485$
DeepSeek-V3.2	2.08	0.50	+2.631	$-0.862 \rightarrow -0.861$

under metadata-only input. Finally, excluding the four clearly high-risk papers in that subset leaves the main reviewer-style ordering and residual-compression pattern unchanged. Together, these checks support the narrower claim that the main results are not driven by a small number of recognizable public manuscripts. A compact summary appears in Table 4, with full audit prompts, recognition criteria, artifact pointers, and slice-by-slice details deferred to the supplementary material.

Input-representation sensitivity. DeepSeek-V3.2 was evaluated through a standardized extracted-text manuscript pipeline because the available endpoint did not support native PDF upload. On a 50-paper subset, PDF-vs-text checks for models supporting both inputs show small shifts for GPT and Claude but larger shifts for Gemini. We therefore treat input representation as a measured robustness dimension, and interpret DeepSeek as reviewer behavior under the fixed extracted-text pipeline.

Within-slice normalized anchors. Because the benchmark spans multiple venue-year slices, we also recomputed model–human alignment under within-slice normalized anchors. The main ordering remains intact: Claude is still best aligned to the human anchor, followed by GPT, while

Gemini and DeepSeek remain substantially less aligned. Under venue-slice z-score normalization, the MAE values are 0.678 for Claude, 0.787 for GPT, 0.817 for Gemini, and 0.811 for DeepSeek; the same ordering also appears under within-slice percentile alignment. Excluding the entire ICLR 2026 slice likewise preserves the same reviewer-profile ordering. This makes it unlikely that the core reviewer-style pattern is an artifact of raw cross-slice score scale alone.

5 Discussion

The main empirical point is not that one model is simply "best," but that current frontier models behave differently as reviewers. Claude is closest to the human score anchor and is also the most stable, though closest to the anchor is not the same as universally "best reviewer." GPT is more permissive on average but preserves many borderline decisions, yielding a decision-boundary hedging pattern rather than a simple score shift. Gemini is substantially more permissive and far less stable under rerun, so its generosity is entangled with broader instability. DeepSeek is by far the most acceptance-oriented model and produces the strongest compression of human score variation. These are not interchangeable instances of a single "LLM reviewer."

The supporting stress tests on Nature Communications and F1000 show that reviewer behavior is context-sensitive and that external anchors can be weak or mismatched. We treat OpenReview as the core scientific setting of the paper and use the external datasets mainly to probe ecological robustness. The operational implication is direct: if LLMs are used in scientific-review workflows, they should be treated as model-specific evaluators with distinct calibration profiles and failure modes, not as swappable generic reviewers.

Artifact / Reproducibility. We provide an anonymous OSF review artifact and will replace it with a persistent public archive after the anonymity period. The review artifact contains `paper_ids.csv` and benchmark manifests with OpenReview IDs, venue-year slices, human scores, and subset membership flags; `prompts/` with P0, P1, title-only, and metadata-only probes; `outputs/` with raw and normalized JSON outputs for all four models; `scripts/` for every reported table and figure; a `README.md` describing how to rebuild the public-source benchmark; and `manifest.json` with file hashes and table/figure provenance. The appendix lists the release objects and protocol details so that the main benchmark, robustness checks, contamination probes, and derived decision-level analyses can be independently audited.

6 Conclusion

Under a shared review protocol and a shared human anchor, frontier LLMs differ in score bias, verdict usage, residual compression, and rerun stability. The strongest evidence comes from the human-anchored OpenReview benchmark. The main result is straightforward: there is no single generic "LLM reviewer." Different models produce measurably different reviewer profiles, and those differences matter both scientifically and practically.

7 Limitations

The benchmark is centered on public OpenReview submissions, so it measures human-anchored reviewer calibration in a realistic public-manuscript setting rather than private, contamination-free scientific judgment. The model set is limited to four frontier systems and a fixed review protocol, so the exact calibration profiles should be remeasured as models and prompts change. Provider interfaces are also not identical: DeepSeek-V3.2 was evaluated through a standardized extracted-text pipeline because the available endpoint did not support native PDF upload, and our PDF-vs-text checks show that input representation can matter for some models.

8 Ethical considerations

This study evaluates LLM behavior in scientific peer review and should not be read as recommending automated reviewer replacement. The results show that model choice can shift scores and verdict boundaries, so LLM-assisted reviewing would require human oversight, calibration checks, and clear disclosure in any deployment. The benchmark uses public manuscripts and public review metadata where available; model inputs exclude reviewer comments, rebuttals, author identities, affiliations, and decision metadata. We release only anonymized reconstruction metadata, prompts, normalized outputs, and analysis scripts for auditability, and we avoid private or confidential manuscripts.

References

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15682, Suzhou, China. Association for Computational Linguistics.

Wei-Lin Chen, Zhepei Wei, Xinyu Zhu, Shi Feng, and Yu Meng. 2025. Do LLM evaluators prefer themselves for a reason? *Preprint*, arXiv:2504.03846.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *Preprint*, arXiv:2402.10886.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

Shomik Jain, Charlotte Park, Matt Viana, Ashia Wilson, and Dana Calacci. 2026. Interaction context often increases sycophancy in LLMs. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*.

Avneet Kaur. 2025. Echoes of agreement: Argument driven sycophancy in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22803–22812, Suzhou, China. Association for Computational Linguistics.

Sung Won Kim and Daniel Khashabi. 2025. Challenging the evaluator: LLM sycophancy under user rebuttal. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22461–22478, Suzhou, China. Association for Computational Linguistics.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel McFarland, and James Zou. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8).

Ryan Liu and Nihar B. Shah. 2023. ReviewerGPT? an exploratory study on using large language models for paper reviewing. *Preprint*, arXiv:2306.00622.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

Giuseppe Russo, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. 2025. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2026. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8:326–336.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. 2025. An empirical analysis of uncertainty in large language model evaluations. In *International Conference on Learning Representations*.

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. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *International Conference on Learning Representations*.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, and Bowen Zhou. 2025. ReviewRL: Towards automated scientific review with RL. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16931–16943, Suzhou, China. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating

large language models at evaluating instruction following. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Preprint*, arXiv:2306.05685. NeurIPS 2023 Datasets and Benchmarks Track.

A Prompt and API Details

This supplement records the structured review prompt used in the main experiments together with the lightly paraphrased variant used in the prompt-robustness check. The task, score anchors, and JSON schema are held fixed across the two versions.

A.1 Main prompt (P0)

```
# Role
You are a senior academic reviewer. Your
↪  task is to provide a rigorous,
objective, and evidence-based peer
↪  review.

# Task Instructions
1. Evaluate the submission ONLY based on
↪  the provided text/PDF.
2. Blind Review Protocol: Disregard any
↪  information about authors,
affiliations, or prior publication
↪  status.
3. No External Search: Do not use web
↪  search or external memory.
4. Scoring Calibration:
   - 1-3: Major flaws, reject.
   - 4-5: Borderline submission with
    ↪  meaningful strengths but notable
      limitations.
   - 6-8: Strong contribution, accept.
   - 9-10: Exceptional contribution.

# Review Process (Step-by-Step)
Step 1: Analyze the technical soundness
↪  and the validity of experimental
results.
Step 2: Assess the novelty relative to
↪  existing literature.
Step 3: Identify specific weaknesses or
↪  missing controls.
Step 4: Generate the final scores based
↪  on the evidence above.

# Output Format
Return a valid JSON object with the
↪  following fields:
{
  "evidence_rationale": "A detailed
  ↪  analytical paragraph focusing on
  technical rigor and logic.",
  "summary": "Short summary of the
  ↪  paper.",
  "soundness": 1-4,
  "presentation": 1-4,
  "contribution": 1-4,
  "strengths": ["list of strings"],
  "weaknesses": ["list of strings"],
  "rating": 1-10,
  "confidence": 1-5,
  "verdict": "reject/borderline/accept"
}
```

A.2 Paraphrased prompt for robustness check (P1)

```
# Role
You are an experienced academic peer
↪  reviewer. Provide a careful,
evidence-based review grounded only in
↪  the manuscript you are given.

# Instructions
1. Judge the paper only from the
↪  provided manuscript content.
2. Ignore any author names,
↪  affiliations, venue cues, or
↪  publication-status
cues if they appear.
3. Do not use outside knowledge,
↪  external search, or remembered
↪  references to
fill gaps.
4. Use the following score anchors:
   - 1-3: serious weaknesses; reject.
   - 4-5: mixed or borderline; promising
    ↪  but with important limitations.
   - 6-8: strong paper; accept.
   - 9-10: outstanding contribution.

# Review Procedure
First assess technical soundness and
↪  whether the empirical evidence
↪  supports
the claims.
Then assess novelty at a high level
↪  based only on what the manuscript
↪  itself
presents.
Then identify concrete weaknesses,
↪  missing controls, or overclaims.
Finally assign the scores and verdict so
↪  that they match the evidence you
discussed.

# Output
Return one valid JSON object with
↪  exactly these fields:
{
  "evidence_rationale": "A concise
  ↪  analytical paragraph focused on
  ↪  rigor,
  evidence, and logic.",
  "summary": "Short summary of the
  ↪  paper.",
  "soundness": 1-4,
  "presentation": 1-4,
  "contribution": 1-4,
  "strengths": ["list of strings"],
  "weaknesses": ["list of strings"],
  "rating": 1-10,
  "confidence": 1-5,
  "verdict": "reject/borderline/accept"
}
```

A.3 Access details

The frontier-model experiments were executed through public provider APIs available during February–March 2026. The primary model identifiers were `gpt-5.4`, `claude-sonnet-4-6`, `gemini-3.1-pro-preview`, and `deepseek-v3.2`. Gemini was accessed through a preview API at the time of study and is therefore interpreted more cautiously in the prompt-robustness and repeatability analyses. Across the main evaluation scripts, decoding

was run at temperature 0, with a nominal output cap of 8192 tokens and request timeout of 900 seconds. Malformed outputs were retried and normalized into the same schema containing summary, strengths, weaknesses, soundness, presentation, contribution, rating, confidence, and verdict. We rename the rationale field to `evidence_rationale` in this supplement for presentation clarity; the literal field name used in the original experimental JSON schema is preserved in the raw logs for exact reproducibility.

A.4 Artifact and release manifest

To make the study auditable rather than merely descriptive, we provide an anonymous OSF review artifact and will replace it with a persistent public archive after the anonymity period. The repository is organized as follows:

- `paper_ids.csv` and benchmark manifests: OpenReview ID, venue-year slice, human scores, and flags for inclusion in the rerun, prompt-robustness, contamination-audit, text-only, and normalized-anchor subsets.
- `prompts/`: the main prompt P0, paraphrased prompt P1, title-only recognizability probe, metadata-only prompt, and schema constraints.
- `outputs/`: raw provider JSON when redistributable, normalized JSON for GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, and DeepSeek-V3.2, plus parse-status logs.
- `scripts/`: analysis scripts that compute each reported table and figure from the manifests and normalized outputs.
- `README.md`: a one-command reconstruction guide describing how to rebuild the benchmark from public sources and rerun the analyses.
- `manifest.json`: hashes for all released files and provenance entries mapping every table and figure in the paper to its input files and scripts.

The prompt text and contamination-probe schemas are recorded verbatim here so that the released outputs can be traced back to exact input specifications.

Table 5: Protocol summary for the main benchmark. The objective is to make provider-access details explicit without turning provider-interface differences into the main scientific claim.

Component	Specification
Primary model IDs	`gpt-5.4`, `claude-sonnet-4-6`, `gemini-3.1-pro-preview`, `deepseek-v3.2`
Access window	February–March 2026
Decoding	Temperature 0 throughout; top-p fixed to 1 where the provider API exposed it and otherwise left at the provider default under deterministic decoding
Nominal output cap	8192 tokens in the main review scripts; 4096 tokens in the contamination-recognition probe
Timeout	900 seconds per request
Main input path	GPT / Claude / Gemini: provider-supported manuscript input with shared review schema; DeepSeek: standardized extracted-text manuscript pipeline
Supplementary text-only checks	A 50-paper all-model extracted-text subset and a 50-paper PDF-versus-extracted-text sensitivity check for the models that supported both representations
Retry policy	Malformed or non-JSON outputs were retried and then normalized into the shared schema when parsing succeeded
Parse/failure handling	Final tables are based on successfully normalized outputs; unresolved parse failures were excluded rather than silently imputed

B Additional Robustness Checks

B.1 Prompt robustness

The main paper visualizes the same prompt-robustness pattern in its corresponding robustness figure.

B.2 Repeatability visualization

The main paper reports repeatability numerically in the text. Figure 5 visualizes the same 34-paper rerun summary: Claude exhibits the lowest rating variability, GPT and DeepSeek are also stable, and Gemini is both more variable in rating and less consistent in verdicts.

B.3 Verdict-boundary behavior across human-score buckets

The main paper notes that GPT-5.4 is not merely shifted upward in average score; it also preserves borderline recommendations much more aggressively across weak, mid-range, and strong papers. Table 7 makes this pattern explicit. GPT assigns borderline to 78.3% of low-scored papers, 59.4% of mid-scored papers, and 32.1% of high-scored papers, whereas Gemini and DeepSeek issue far fewer borderline verdicts and collapse more strongly toward decisive accept/reject outcomes. The main paper also visualizes the corresponding score distributions.

Table 6: Prompt-robustness summary on the 30-paper stratified OpenReview subset. Negative MAE change means that the paraphrased prompt moved the model closer to the human mean. Higher same-verdict ratio is better.

Model	MAE change (toward 0)	Same-verdict ratio ↑	Interpretation
GPT-5.4	-0.021	0.967	highly stable
Claude 4.6	+0.078	0.933	highly stable
Gemini 3.1 Pro	-0.543	0.633	confounded by rerun instability
DeepSeek-V3.2	-0.650	0.933	prompt-sensitive but verdict-stable

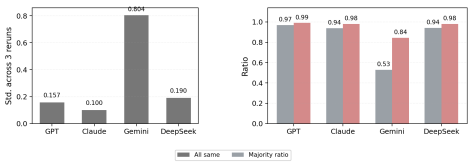

Figure 5: Repeatability on the 34-paper rerun subset. Left: rating standard deviation across three reruns. Right: exact all-same verdict ratio (gray) and majority-verdict consistency (red).

B.4 Decision-level calibration and simulated reviewer assignment

To make reviewer style concrete at the decision level, we map the paper-level human mean into coarse verdict buckets using $< 4 \mapsto$ reject, 4–6 $\mapsto$ borderline, and $\geq 6 \mapsto$ accept. Table 8 reports macro-F1, balanced accuracy, weighted κ, and per-bucket recall under that mapping. This confirms that Claude is the most balanced overall, GPT best preserves the borderline region, and Gemini/DeepSeek heavily favor accept-like decisions. Table 9 then asks a deployment-style question: if one LLM score is added as an extra reviewer to the existing human scores, how often does the final coarse bucket move? DeepSeek changes the decision boundary most often, Claude the least. Finally, Table 10 places the models against the observed human-reviewer envelope rather than just the human mean.

B.5 Input-representation sensitivity

At the time of our experiments, the DeepSeek-V3.2 endpoint available through the provider's official chat interface did not support a native manuscript/PDF attachment path directly comparable to the other providers. We therefore evaluated DeepSeek under a standardized extracted-text manuscript pipeline. To gauge whether this asymmetry could plausibly dominate the main findings, we ran an input-representation sensitivity check on the models that supported both PDF and extracted-text evaluation.

Table 7: Verdict usage across human-score buckets on the 250-paper OpenReview benchmark. Borderline rate isolates each model's tendency to preserve middle verdicts rather than issuing decisive reject/accept recommendations.

Bucket	Model	n	Reject	Borderline	Accept	Borderline rate ↑
Low (< 4)	GPT-5.4	69	9	54	6	0.783
Low (< 4)	Claude 4.6	69	29	33	7	0.478
Low (< 4)	Gemini 3.1 Pro	69	32	4	33	0.058
Low (< 4)	DeepSeek-V3.2	69	2	16	51	0.232
Mid (4–6)	GPT-5.4	128	0	76	52	0.594
Mid (4–6)	Claude 4.6	128	3	63	62	0.492
Mid (4–6)	Gemini 3.1 Pro	128	16	9	103	0.070
Mid (4–6)	DeepSeek-V3.2	128	0	4	124	0.031
High (≥ 6)	GPT-5.4	53	0	17	36	0.321
High (≥ 6)	Claude 4.6	53	0	13	40	0.245
High (≥ 6)	Gemini 3.1 Pro	53	4	1	48	0.019
High (≥ 6)	DeepSeek-V3.2	53	0	0	53	0.000

Table 8: Decision-level calibration on the 250-paper OpenReview benchmark under the coarse human-anchor mapping $< 4 \mapsto$ reject, 4–6 $\mapsto$ borderline, and $\geq 6 \mapsto$ accept. Higher macro-F1, balanced accuracy, weighted κ, and bucket recall are better.

Model	Macro-F1 ↑	Bal. acc. ↑	QWK ↑	Borderline recall ↑	Accept recall ↑	Reject recall ↑
GPT-5.4	0.424	0.468	0.377	0.594	0.679	0.130
Claude 4.6	0.533	0.556	0.491	0.492	0.755	0.420
Gemini 3.1 Pro	0.354	0.480	0.274	0.070	0.906	0.464
DeepSeek-V3.2	0.163	0.353	0.097	0.031	1.000	0.029

B.6 Public-manuscript contamination audit

We do not claim that the OpenReview benchmark is contamination-free. Instead, we report three minimal controls. First, a full-benchmark title-only recognition probe using Claude Sonnet 4.6 as a strong recognizability audit marks only 7 of 250 papers as recognizable. Excluding those 7 papers from the full benchmark leaves the overall ordering essentially unchanged (mean deltas: GPT $+0.840$, Claude $+0.347$, Gemini $+1.437$, DeepSeek $+2.676$). Second, on a 50-paper multi-model subset, a metadata-only review condition (title plus abstract) does not recover the same behavior as full-manuscript evaluation, especially for DeepSeek-V3.2. Third, excluding the four clearly high-risk papers in that subset leaves the model ordering and residual-compression pattern unchanged.

Table 9: Counterfactual extra-reviewer simulation. For each paper, we add one LLM score to the existing set of human reviewer scores and recompute the coarse human-anchor bucket. Higher flip rates indicate larger decision-boundary impact.

Model	Boundary-flip rate ↓	Reject→nonreject ↓	Low→borderline/accept ↓	Borderline→accept ↓
GPT-5.4	0.168	0.060	0.060	0.072
Claude 4.6	0.104	0.044	0.044	0.024
Gemini 3.1 Pro	0.232	0.092	0.092	0.112
DeepSeek-V3.2	0.284	0.152	0.152	0.132

Table 10: Human-disagreement envelope positioning. In-range rate is the fraction of papers on which the model score falls within the observed human-reviewer score range for that paper. Higher in-range rate and lower MAE are better.

Model	In-range rate ↑	Above-range rate ↓	Below-range rate ↓	MAE vs human mean ↓	MAE vs leave-one-out ↓
GPT-5.4	0.736	0.232	0.032	1.172	1.196
Claude 4.6	0.780	0.148	0.072	0.898	0.950
Gemini 3.1 Pro	0.432	0.488	0.080	1.922	1.941
DeepSeek-V3.2	0.308	0.692	0.000	2.655	2.652

Table 11: Full reviewer-style fingerprint corresponding to the compact summary in the main paper. Larger positive bias means stronger upward scoring relative to the human anchor. Lower low-score accept rate, weaker residual compression (slope closer to 0), lower rerun std, higher prompt verdict agreement, lower counterfactual boundary-flip rate, and higher in-range rate indicate a more conservative and stable reviewer profile.

Model	Bias Δ	Borderline rate ↑	Low-score accept ↓	Residual slope (toward 0)	Rerun std ↓	Prompt agr. ↑	Flip rate ↓	In-range rate ↑
GPT-5.4	+0.833	0.588	0.087	-0.450	0.157	0.967	0.168	0.736
Claude 4.6	+0.349	0.436	0.101	-0.252	0.100	0.933	0.104	0.780
Gemini 3.1 Pro	+1.445	0.056	0.478	-0.277	0.804	0.643	0.232	0.432
DeepSeek-V3.2	+2.645	0.080	0.739	-0.655	0.190	0.933	0.284	0.308

Table 12: Input-representation sensitivity on a 50-paper OpenReview subset. Lower absolute score difference and higher verdict agreement indicate greater robustness to alternative public input representations.

Model	Mean absolute score difference ↓	Same-verdict ratio ↑
GPT-5.4	0.34	0.90
Claude 4.6	0.40	0.74
Gemini 3.1 Pro	1.30	0.66

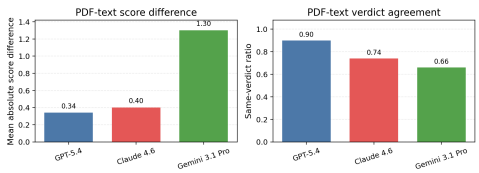

Figure 6: Input-representation sensitivity on a 50-paper OpenReview subset. Left: mean absolute score difference between PDF and extracted-text evaluation (lower is better). Right: same-verdict ratio (higher is better).

Table 13: Operational definition of the title-only recognition probe. We use this audit to bound obvious public-manuscript recognizability rather than to prove the absence of latent contamination.

Component	Specification
Input	Manuscript title only
Probe instruction	Ask whether the model recognizes the paper well enough to infer probable venue/platform, likely acceptance outcome, or specific reviewer concerns
Output schema	JSON fields for `recognized_paper`, guessed venue, guessed decision, guessed reviewer concerns, confidence, and short rationale
Recognized criterion	Counted as recognizable only when the parsed output explicitly marks `recognized_paper=true`; generic topic guesses without that flag are not counted as recognized
Audit model	Claude Sonnet 4.6, used as a strong recognizability probe
Recognized count	7 of 250 papers
Exclusion check	Recomputing the benchmark after removing those 7 papers leaves the model ordering unchanged

Table 14: Full-benchmark title-only recognition audit with Claude Sonnet 4.6. Even under this strong recognizability probe, only 7 of 250 papers are marked recognizable from title alone. Lower recognition is preferable.

Model used for audit	n	Recognized ↓	Recognition rate ↓
Claude 4.6	250	7	0.028

Table 15: Four-model title-only recognition audit on the 50-paper subset. The recognized counts remain low across all four models, which makes the contamination-risk signal less dependent on a single audit model. Lower recognition is preferable.

Model	n	Recognized ↓	Recognition rate ↓
GPT-5.4	50	1	0.02
Claude 4.6	50	4	0.08
Gemini 3.1 Pro	50	1	0.02
DeepSeek-V3.2	50	0	0.00

Table 16: Title-plus-abstract-only control on a 50-paper OpenReview subset. Larger differences from full-manuscript behavior indicate that title-plus-abstract metadata is not sufficient to reproduce the full-manuscript pattern. Lower MAE diff and higher same-verdict ratio indicate closer metadata/full agreement.

Model	n	Full mean	Metadata mean	MAE diff ↓	Same-verdict ratio ↑
GPT-5.4	50	5.78	6.18	0.76	0.68
Claude 4.6	50	5.50	5.70	0.56	0.78
Gemini 3.1 Pro	50	6.06	6.90	1.36	0.70
DeepSeek-V3.2	50	7.70	5.62	2.08	0.50

Table 17: Within-slice normalized-anchor robustness on the 250-paper OpenReview benchmark. Scores are normalized separately within each venue-year slice before model–human comparison. Lower MAE/RMSE and higher Spearman are better.

Model	n	MAE (z) ↓	RMSE (z) ↓	MAE (percentile) ↓	Spearman (z) ↑
GPT-5.4	250	0.787	0.982	0.217	0.500
Claude 4.6	250	0.678	0.857	0.196	0.596
Gemini 3.1 Pro	250	0.817	1.025	0.239	0.441
DeepSeek-V3.2	250	0.811	1.033	0.228	0.384

Table 18: All-model text-only comparison on the same 50-paper OpenReview subset. The reviewer-style ordering remains unchanged under a standardized extracted-text representation. Larger positive mean delta indicates stronger upward bias relative to the human anchor.

Model	n	Human mean	Model mean ↑	Mean delta
Claude 4.6	50	5.069	5.54	+0.471
GPT-5.4	50	5.069	5.80	+0.731
Gemini 3.1 Pro	50	5.069	6.50	+1.431
DeepSeek-V3.2	50	5.069	7.70	+2.631

Table 19: High-risk exclusion audit on the 50-paper OpenReview subset. Removing the four title-recognizable papers does not change the overall ordering. Residual slopes closer to zero indicate weaker compression.

Model	n	Mean delta	Residual slope (toward 0)
Claude 4.6	46	+0.415	-0.548
GPT-5.4	46	+0.697	-0.512
Gemini 3.1 Pro	46	+0.980	-0.485
DeepSeek-V3.2	46	+2.632	-0.861

C Cross-Platform Stress Tests

C.1 Nature Communications high-quality published-paper set

As a supporting stress test on strong papers, we compare OpenReview papers with human mean score at least 6 ($n = 53$) against a high-quality published-paper set from Nature Communications ($n = 49$). All four models become more acceptance-oriented on the published set. GPT and Claude still preserve some boundary sensitivity, while Gemini and DeepSeek collapse much more strongly toward accept-only behavior. We interpret this comparison as a context-sensitive high-quality stress test rather than as a score-anchored benchmark.

C.2 F1000 public-review-state benchmark

Unlike Nature Communications, F1000 exposes public reviewer outcomes. After removing 10 severely damaged redaction cases, we retain 43 papers and map public reviewer states into reject-like, borderline-like, and accept-like levels. Agreement with these coarse public-review anchors is weak overall, so we treat F1000 as a diagnostic benchmark rather than as a strong validation set.

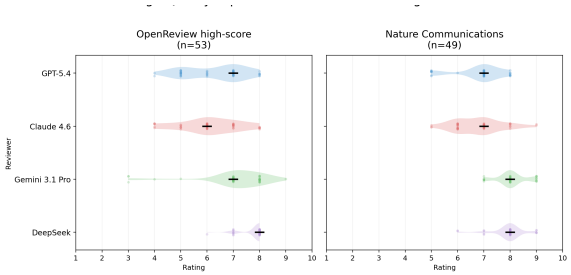

Figure 7: Model rating distributions on OpenReview high-score papers and Nature Communications published papers.

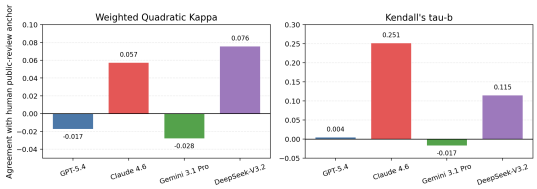

Figure 8: Ordinal agreement on F1000 under public human-review-state anchors.

D Qualitative Case Studies

Table 20 summarizes eight illustrative Nature Communications cases. GPT-5.4 and Claude Sonnet 4.6 most often overlap with human concerns at the level of mechanism, controls, and stability, while Gemini and DeepSeek more often produce higher-level and more permissive critiques.

Table 20: Illustrative Nature Communications case studies. GPT and Claude most often recover the same technical concern types as human reviewers, while Gemini and DeepSeek more often remain higher-level.

ID	Domain	Human main concerns	GPT-5.4	Claude 4.6	Gemini 3.1	DeepSeek
NC-001	Hydrogel sensor	Mechanism controls, lipid migration evidence, longer-term sensing, friction stability	Strong mechanism overlap	Strong overlap; adds practical concerns	Partial overlap; more generic	Partial overlap; high-level only
NC-002	Core-sheath hydrogel	Mechanical validation, SAXS interpretation, benchmark framing	Mechanical validation focus	Strongest on rigor and artifact risk	Looser; mostly application limits	Generic stability concerns
NC-003	Dynamic hydrogel	Mechanism and predictive theory	Strong overlap on mechanism and controls	Strongest on predictive framework	Moderate overlap, more positive	Moderate overlap, high-level only
NC-006	MoS_2 hydrogel	Mechanical characterization, universality, mechanism	Strong overlap on mechanism	Strong overlap; more skeptical on core claim	Partial overlap, milder	Partial overlap, more generic
NC-011	LIG sensing	Decoupling evidence, presentation clarity, mechanism	Strong overlap on decoupling and stability	Strong overlap plus benchmarking concerns	Partial overlap, more generous	Partial overlap, higher-level concerns
NC-012	TENG field emission	Reproducibility, environment dependence, mechanism validation	Strong overlap; critical on direct evidence	Strong overlap; most skeptical on proof	Partial overlap, more positive	Partial overlap, generous
NC-013	DC-TENG	Mechanical evidence, practical demonstration, theory clarity	Strong overlap on mechanism and theory	Strong overlap on theory and benchmarking	Partial overlap, application-oriented	Partial overlap, broader comments
NC-014	Electrochemistry	Side reactions, mechanism attribution, related work	Strong overlap on carbon balance and mechanism	Strong overlap; strongest on process accounting	Partial overlap, scale-up oriented	Partial overlap, high-level process concerns

E Dataset Construction Details

E.1 OpenReview benchmark construction

The OpenReview benchmark was assembled from submissions with recoverable manuscript files and usable reviewer-score metadata. The retained 250-paper set spans ICLR 2024 (76 papers), NeurIPS 2024 (76 papers), ICLR 2025 (75 papers), and ICLR 2026 (23 papers), and covers the full observed human-score range in our sample, from 0 to 8. For every retained paper, we required a usable manuscript, at least one numeric reviewer rating, and sufficient metadata to align the paper ID, manuscript file, and review record.

A venue-stratified robustness check preserves the same model ordering in every slice. The model–human mean deltas are: ICLR 2024 = GPT +0.847, Claude +0.189, Gemini +1.320, DeepSeek +2.702; NeurIPS 2024 = GPT +0.692, Claude +0.416, Gemini +1.297, DeepSeek +2.153; ICLR 2025 = GPT +0.709, Claude +0.256, Gemini +1.456, DeepSeek +2.589; ICLR 2026 = GPT +1.659, Claude +0.964, Gemini +2.312, DeepSeek +4.268. Excluding the entire ICLR 2026 slice leaves the main reviewer-style ordering unchanged.

E.2 Nature Communications curation

The Nature Communications external-validation set contains 49 papers with publicly accessible article PDFs and corresponding peer-review files.

Table 21: Observed OpenReview score ranges in the retained benchmark. We use the native numeric reviewer scores exposed on OpenReview and do not rescale them across venue-year slices. The paper-level mean range is narrower than the raw reviewer-score range because it averages multiple reviews.

Slice	Papers	Raw min	Raw max	Mean min	Mean max
ICLR 2024	76	1.0	8.0	2.5	8.0
NeurIPS 2024	76	1.0	9.0	2.25	7.33
ICLR 2025	75	1.0	8.0	2.0	6.5
ICLR 2026	23	0.0	10.0	0.0	5.5

We use these papers as a high-quality published-paper set rather than as a score-anchored benchmark.

E.3 F1000 curation

The F1000 external-validation set was constructed from redacted article PDFs paired with public reviewer outcomes. We manually excluded the 10 most severely damaged redaction cases and retained 43 papers. Public reviewer states were mapped into a three-level ordinal anchor: not-approved $\rightarrow$ 0, approved-with-reservations $\rightarrow$ 1, and approved $\rightarrow$ 2.