When Can an AI Reviewer Replace a Human? Human-Equivalence Selective Review

Anonymous authors
Paper under double-blind review

Abstract

As AI systems spread through the research pipeline, peer review faces pressure to scale without losing human judgment; generating a plausible critique is different from reproducing how human reviewers decide. We study 454 ICLR submissions from 2024–2026 with 1,810 public human reviews and evaluations from four foundation models. The sharpest reliability gap occurs in the half-level region around the human recommendation boundary: moving outside this region improves AI–human agreement, although rare extreme scores are not monotonically safer. This observation yields **Human-Equivalence Selective Review (HESR)**, a fixed assignment rule that lets one AI review occupy a review slot when its ordinal score lies at least half a level from the human-aligned boundary and otherwise routes the paper to a person. HESR requires no training, model-specific parameters, batch ranking, or multi-model inference. Across 99 deduplicated papers and 396 model–paper judgments, HESR retains 54.8% of reviews and raises agreement with individual human recommendations from 67.6% to 73.6% (paper-cluster bootstrap 95% CI for the gain, 2.5–9.7 percentage points). Historical decision accuracy increases from 73.5% to 87.1%; all four base models and all six evaluation blocks improve, with 109 accept-side and 108 reject-side assignments. Two frozen tests reproduce the effect on 40 unseen papers. On a quality-controlled cross-year test, the gain remains positive after adversarially removing one human review from every paper. These results show that a single AI reviewer can match the variability of human recommendations on an identifiable subset of papers while preserving human review for the least stable cases.

1 Introduction

AI systems increasingly participate in literature search, experimentation, manuscript production, and scientific critique (Weng et al., 2025). This expanding role makes peer review a natural candidate for automation. Automated feedback can help authors, but conference operation poses a stricter question: **when can an AI review occupy a human review slot without changing the recommendation process?**

Current evaluation often measures whether model critiques are useful, specific, or topically similar to human reviews (Liang et al., 2023; Gao et al., 2024; D'Arcy et al., 2024). These properties matter, yet they do not establish reviewer replacement. A model can identify legitimate weaknesses and still assign a systematically different recommendation because it penalizes evidence gaps, presentation flaws, or missing experiments more severely than human reviewers. Conversely, a single human score is not an error-free target. Reviewers disagree, and repeated committees can reach different decisions on the same papers (Beygelzimer et al., 2023). Replacement should therefore be evaluated against the distribution of human recommendations, on the same papers, rather than against one presumed ground truth.

We compare complete-paper reviews from Qwen3.8-Max, GLM5.3, GPT-5.6-Sol, and DeepSeek DS4Pro with public ICLR reviews. The central empirical pattern is simple: AI reliability depends on the position of its score relative to the human decision boundary. The half-level boundary region has the lowest AI–human agreement; agreement rises immediately outside it, although the small set of extreme scores is not monotonically more reliable. The same fixed exclusion region transfers across model families and both recommendation directions. More elaborate alternatives, including

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

learned probability calibration, model-specific mappings, and repeated sampling from one model, fail to transfer as consistently.

This result motivates **Human-Equivalence Selective Review (HESR)**. A model reads one paper, produces one review and one ordinal recommendation, and HESR applies a fixed rule. Scores at least half a level from the human-aligned boundary may occupy one review slot; boundary cases return to human review. Four models are used only to test transfer. Their outputs are never combined at deployment.

HESR turns reviewer replacement into an assignment problem. The method does not require every AI review to be human-equivalent. It identifies the cases in which a single model already behaves like another draw from the human reviewer population. This formulation produces three contributions:

1. **A reviewer-seat criterion.** We evaluate AI recommendations against individual human judgments and compare them with human–human agreement on the same selected papers.
2. **A fixed single-model method.** HESR uses one interpretable ordinal margin, without training, model-specific tuning, batch-dependent top-k selection, or inference-time voting.
3. **Cross-model and cross-year evidence.** Across 99 papers and 396 judgments, the same rule improves all four model families and both prediction directions. Two frozen confirmation batches reproduce the gain on 40 unseen papers.

2 Related work

Automated scientific review. PeerRead and NLPeer support computational research on review generation, rating prediction, and acceptance decisions (Kang et al., 2018; Dycke et al., 2023). Large-scale studies show that model feedback can overlap with human comments and provide useful author feedback (Liang et al., 2023). Reviewer2 models the aspects covered by a review, while MARG coordinates specialized agents to broaden critique coverage (Gao et al., 2024; D'Arcy et al., 2024). ReviewCritique and ReviewerToo examine review deficiencies and conference-style evaluation (Du et al., 2024; Sahu et al., 2025). These systems improve the content of a generated review. We study the downstream assignment decision: whether its recommendation can replace one human judgment.

Differences between human and AI reviewers. Recent benchmarks compare scores, critique topics, and prompt-injection susceptibility (Li et al., 2026a). Rhetorical edits can change AI ratings while leaving the scientific evidence fixed (Li et al., 2026b), and statistical bridge models can narrow systematic gaps between human and model judgments (Polo et al., 2025). Our experiments likewise find scale and preference differences, but the most transferable signal is the score's distance from the human decision boundary. HESR uses that signal to decide when automation is appropriate.

Evaluation without a single gold standard. GEM and GRE-bench evaluate generated judgments and scientific reviews without assuming that one reference is definitive (Xu et al., 2025). Our target is narrower and operational: whether one AI recommendation can occupy a human reviewer seat. We therefore preserve the distribution of individual recommendations and compare the selected AI judgments with human–human agreement on the same papers.

Calibration and selective prediction. Post-hoc calibration aligns confidence with observed outcomes without retraining a base model (Guo et al., 2017). Selective prediction adds a reject option and evaluates the resulting risk–coverage trade-off (Geifman & El-Yaniv, 2019). Trust or Escalate applies calibrated selective evaluation to pairwise LLM judging and estimates confidence through simulated annotators (Jung et al., 2025). HESR addresses full-paper peer review, where the observable decision is an ordinal recommendation and extra simulated reviewers are costly. Its contribution is a reviewer-seat target and a transferable confidence proxy: the fixed boundary margin is evaluated against individual human recommendations, human–human variability on the same assignments, four independent reviewer models, and frozen in-year and cross-year tests.

Human disagreement. Peer-review outcomes include irreducible variation across reviewers and committees (Beygelzimer et al., 2023). We therefore distinguish agreement with individual human recommendations, agreement between humans on the same subset, and accuracy against the histor-

ical conference decision. The first is the primary replacement outcome; the other two show whether selection exploits unusually easy or one-sided cases.

3 PROBLEM SETTING AND DATA

For paper i, human reviewer j assigns an ordinal recommendation h_{ij} and AI model m assigns r_{im}. We harmonize recent ICLR rating forms onto six semantic levels, $0, \ldots, 5$, with level 3 as the accept/reject boundary. Let a_i be the fraction of retained human reviewers recommending acceptance. If the AI predicts acceptance, its agreement on paper i is a_i; if it predicts rejection, agreement is $1 - a_i$. We average first within each paper and then across papers, so papers with more archived reviews do not receive greater weight.

Human–human agreement is the mean pairwise decision agreement among reviewers of the same paper. Historical decision accuracy compares the AI recommendation with the conference outcome. These quantities answer different questions: resemblance to one human reviewer, variability among humans, and consistency with the final decision.

3.1 RESEARCH CORPUS

We use public ICLR papers and OpenReview fields from 2024–2026 (JerMa88, 2024; 2025; 2026). After deduplication by OpenReview paper identifier, the research corpus contains 454 submissions and 1,810 human reviews: 90 papers from 2024, 252 from 2025, and 112 from 2026; 215 were accepted and 239 rejected. This corpus supports problem diagnosis, method design, controlled comparisons, and confirmation experiments.

Table 1: Research corpus after paper-level deduplication.

Study corpus	Papers	Human reviews	ICLR years	Accept / reject
Deduplicated research pool	454	1,810	2024–2026	215 / 239

Model packets contain the abstract and extracted scientific text. Before review, a deterministic sanitation pass removes titles, front matter, acknowledgments, publication-status language, author and venue metadata, links, and stable identifiers. Human reviews, rating fields, and decision fields are withheld. All reported model outputs were generated from these sanitized, outcome-blinded packets. Every model–paper judgment is produced in a fresh context. The four validation models receive the same task definition and output schema and are evaluated separately.

3.2 HESR EVALUATION SETS

The HESR evidence base contains 99 unique papers and 396 model–paper judgments across six blocks. We give the blocks semantic names according to their role: development (32 papers), leave-one-paper-out audit (8), two stabilization replications (12 each), an in-year frozen test (16 unseen ICLR 2025 papers), and a cross-year frozen test (24 previously unused papers from ICLR 2024 and 2025). Paper-level deduplication retains the newest frozen measurement when a paper also appeared in an earlier analysis. The four validation models review every paper independently.

The cross-year frozen test is balanced by year and conference outcome and contains both decisive and contested human-review profiles. It also controls reference quality: every paper has at least three reviews with complete summaries, strengths, weaknesses, and paper-specific content, and exact duplicates count once. Model packets never contain these reviews. The primary reference averages retained reviewer decisions; robustness analyses use the majority direction and remove human reviews one at a time.

Table 2: Evidence blocks used for method construction and frozen validation.

Block	Role	Papers	Jdg.	Use
Develop.	Locate transferable reliability structure	32	128	Rule formation
LOO audit	Check dependence on individual papers	8	32	Internal audit
Stab. I	Compare learned and fixed selectors	12	48	Method check
Stab. II	Test transfer after batch shift	12	48	Method check
In-year	Unseen ICLR 2025 papers	16	64	Frozen validation
Cross-year	Unseen 2024–2025 papers after review-quality screening	24	96	Frozen validation

4 Human-Equivalence Selective Review

4.1 From review-policy learning to a fixed assignment rule

We first tested whether human review behavior could be reproduced by learning richer policies over critique type, model identity, repeated judgments, or estimated human agreement. These alternatives isolate different explanations for AI–human disagreement. Table 3 reports the result that determined the final design: richer policies can fit a batch or improve a secondary score metric, but the component that repeatedly transfers is the paper-level ordinal position relative to the recommendation boundary.

Table 3: Candidate method families and the evidence used to select the final HESR rule.

Design	Question	Evidence	HESR consequence
Issue weighting	Can critique categories recover human preferences?	Broader discovery produced no stable recommendation rule under matched budgets.	No defect taxonomy.
Learned selector	Can fitted probabilities identify human-like reviews?	Agreement changed 65.6%→81.7% in Stabilization I, then 63.1%→62.9% on the next unseen batch.	Remove fitting and batch top-k.
Provider maps	Does each provider need a separate score scale?	Continuous error sometimes fell, but transfer varied across models and batches.	Use one shared boundary.
Repeat calls	Can resampling recover heterogeneous-review gains?	Four GPT reviews had error correlation 0.990; median aggregation did not improve MAE or rank correlation.	Use one review.
Fixed margin	Is boundary distance a transferable reliability signal?	The half-level rule improves all six blocks and all four base models after deduplication.	Final assignment rule.

This progression is an ablation of method capacity rather than a search for the largest development-set gain. It removes components whose apparent benefit depends on the batch, provider, or additional calls, leaving a rule that can be applied independently to one new paper.

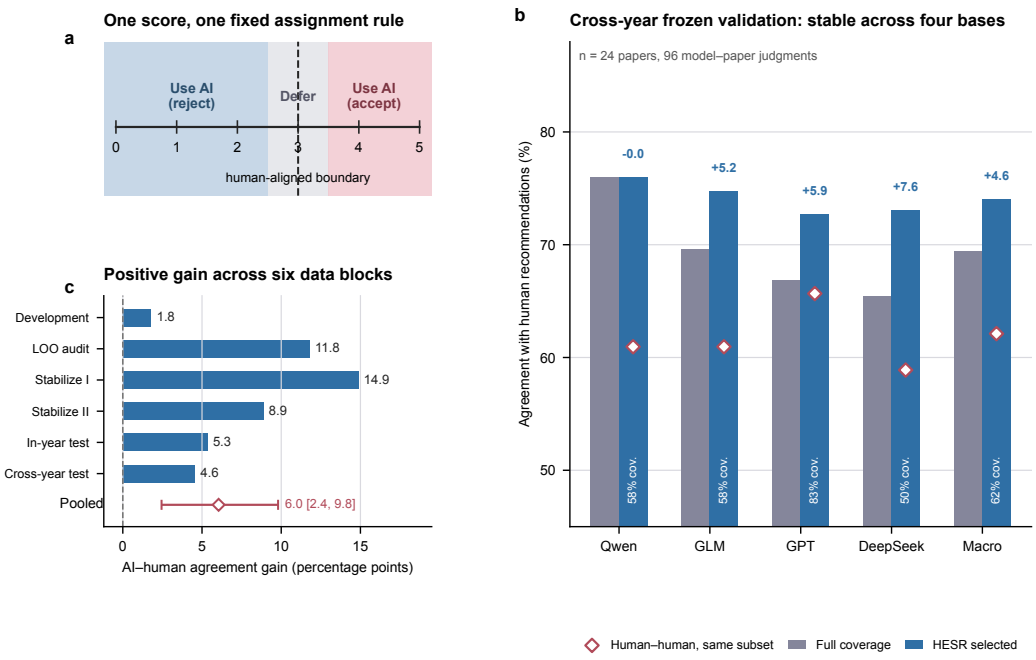

Figure 1: Human-Equivalence Selective Review and its evidence. **a**, A fixed half-level region around the human-aligned boundary routes accept- and reject-side judgments to AI and boundary cases to people. **b**, On the 24-paper cross-year frozen test, HESR improves three base models and leaves Qwen unchanged at reported precision; diamonds show human–human agreement on each selected subset. **c**, All six evidence blocks show positive aggregate gains. The pooled point and interval use 99 deduplicated papers and paper-cluster bootstrap resampling.

4.2 Fixed HESR rule

The diagnostic result is that disagreement concentrates in the half-level region around the human recommendation boundary. HESR converts that observation into one assignment rule:

$$\text{use AI review}_{im} = \mathbb{1}[|r_{im} - 3| \geq 0.5] \,.$$

An AI score of 2.5 or below supplies a reject-side review slot; a score of 3.5 or above supplies an accept-side slot. Scores strictly between 2.5 and 3.5 are routed to a person. The rule reads one model's absolute ordinal score. It does not rank a test batch, estimate a provider-specific threshold, or inspect another model's answer.

HESR returns the review, original ordinal score, direction, and a transparent use/defer decision. A deferred critique may still assist a person, but it does not replace a human recommendation. The margin of 0.5 was fixed before both frozen tests.

A shared monotone score projection, $0.710r + 0.902$, is evaluated as a secondary reporting component. It can reduce distance to mean human scores while preserving paper order, but it does not determine HESR selection or the accept/reject recommendation.

5 Experimental design

The evaluation asks whether the fixed rule improves reviewer-seat reliability, transfers across base models, works in both recommendation directions, and remains stable when the human reference changes.

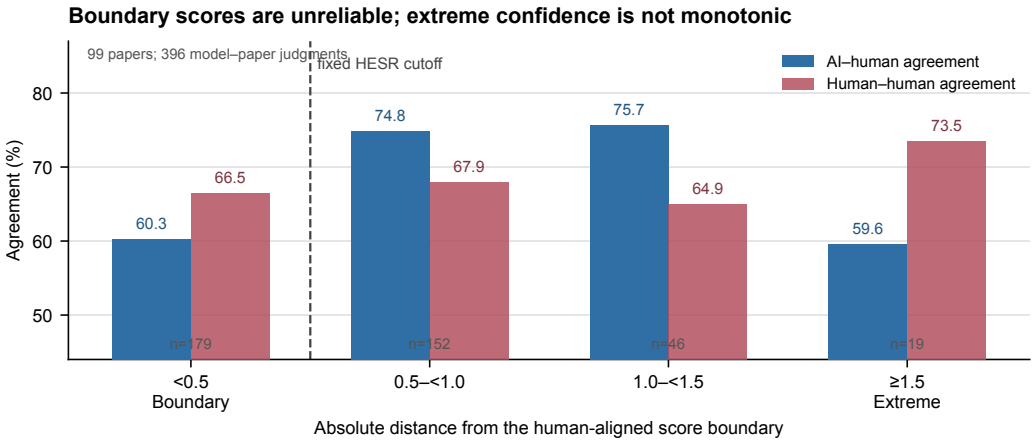

Figure 2: Agreement by absolute ordinal distance from the human-aligned boundary. Bars show AI–human and human–human agreement; labels give model–paper judgment counts. The dashed line marks the fixed HESR cutoff. Agreement rises directly outside the boundary region and then saturates; the rare extreme bin is not more reliable.

The primary metric is the change in AI–human agreement from full coverage to the HESR-selected subset. We also report coverage, human–human agreement on exactly the same subset, historical decision accuracy, and results by model and predicted direction. Confidence intervals use 10,000 paper-cluster bootstrap replicates, keeping all model outputs and human reviews for a sampled paper in one cluster.

The in-year and cross-year frozen tests were executed after fixing the score boundary, half-level margin, output schema, model identities, and analysis scripts. The in-year test contains 16 unseen papers and 64 model judgments. The cross-year test contains 24 unseen papers and 96 judgments. Its success criteria require selected agreement above full coverage, selected agreement no more than five points below human–human agreement, nonnegative gains for at least three models, and no more than a two-point loss in historical decision accuracy.

Controls test the leading alternative explanations. Learned score-to-agreement models and provider-specific maps test whether continuous or model-dependent calibration is required. Repeated GPT reviews test whether ordinary resampling can explain the gain. Direction-stratified results test whether selection merely favors acceptance. Human-reference deletion analyses test whether one low-quality or anomalous review determines the result. The pooled analysis includes every completed four-model block under the unchanged rule and deduplicates papers before estimating the aggregate effect.

6 Results

6.1 The half-level boundary region identifies unreliable judgments

The 179 judgments within half a level of score 3 have 60.3% AI–human agreement, compared with 74.8% for the 152 judgments at distances of 0.5–<1.0 and 75.7% for the 46 judgments at 1.0–<1.5 (Fig. 2). The 19 extreme judgments have 59.6% agreement. Thus the transferable signal is the unreliable boundary region, not a monotonic claim that every larger score magnitude is safer. Human–human agreement ranges from 64.9% to 73.5% across the same bins, so the pattern cannot be explained solely by selecting papers on which humans always agree.

The unchanged half-level rule improves aggregate agreement in all six evidence blocks, with gains from 1.8 to 14.9 percentage points (Fig. 1c). Each judgment uses one review from one base model. In the deduplicated evidence, HESR selects 109 accept-side and 108 reject-side judgments, so the result is not produced by inference-time model pooling or one-sided selection.

6.2 FROZEN TESTS REPRODUCE THE EFFECT

The in-year frozen test selects 38 of 64 judgments (59.4% coverage). AI–human agreement rises from 65.3% to 70.7%, and historical decision accuracy rises from 75.0% to 84.2%. Human–human agreement on the selected papers is 71.1%. All four base models improve.

The cross-year frozen test selects 60 of 96 judgments (62.5% coverage). Agreement rises from 69.4% to 74.0%, a gain of 4.6 percentage points. Historical decision accuracy rises from 81.3% to 93.3%. GLM, GPT, and DeepSeek improve by 5.2, 5.9, and 7.6 points; Qwen changes by less than 0.1 point. The selected set contains 29 accept and 31 reject predictions and passes all four frozen success criteria.

Table 4: Frozen-test and pooled reviewer-seat results.

Set	Papers / judgments	Coverage	Full	HESR	Gain	Decision: full→HESR
In-year frozen	16 / 64	59.4%	65.3%	70.7%	+5.3 pp	75.0%→84.2%
Cross-year frozen	24 / 96	62.5%	69.4%	74.0%	+4.6 pp	81.3%→93.3%
Pooled, deduplicated	99 / 396	54.8%	67.6%	73.6%	+6.0 pp	73.5%→87.1%

Across the two frozen tests, 98 of 160 judgments are selected. Agreement rises from 67.8% to 72.7%; the paper-cluster interval for the 4.9-point gain is [-0.5, 10.9]. Both tests independently reproduce the direction, while the larger pooled analysis yields the narrower interval.

6.3 THE POOLED EFFECT TRANSFERS ACROSS MODELS AND DECISIONS

After paper-level deduplication, HESR retains 217 of 396 judgments, or 54.8% coverage. Agreement increases from 67.6% to 73.6%. The 6.0-point gain has a paper-cluster bootstrap 95% interval of [2.5, 9.7]. Human reviewers agree with one another on 67.7% of judgments for the same selected papers; the AI–human minus human–human interval is [-0.1, 10.9] points. Selected AI recommendations therefore fall within the observed range of human-review variability.

Table 5: Pooled transfer across the four independently evaluated base models.

Base model	Coverage	Full agreement	HESR agreement	Gain
Qwen3.8-Max	54.5%	72.1%	74.9%	+2.8 pp
GLM5.3	46.5%	67.6%	75.7%	+8.1 pp
GPT-5.6-Sol	69.7%	64.9%	71.4%	+6.5 pp
DeepSeek DS4Pro	48.5%	65.8%	73.5%	+7.7 pp

Accept-side agreement gains 11.5 points and reject-side agreement gains 1.9 points. Historical decision accuracy reaches 86.2% for selected accept predictions and 88.0% for selected reject predictions. The same rule therefore transfers across four independent model environments and retains bidirectional coverage, even though its largest gain occurs on the accept side.

6.4 HUMAN-REFERENCE AND METHOD CONTROLS

The cross-year result remains positive under three human-reference constructions. Against the majority human direction, agreement rises from 79.2% to 83.3%. Across 98 single-review-deletion analyses, the gain ranges from 4.0 to 5.7 points. In an adversarial calculation that removes one review from every paper and chooses each deletion to minimize the method effect, the gain remains positive at 0.4 points.

The method controls explain why the fixed rule is retained. Learned human-agreement calibration produces a 16.0-point gain in the first stabilization replication but changes agreement by -0.2 point in the next unseen batch. Four repeated GPT reviews have pairwise error correlation 0.990, and median aggregation does not improve MAE or rank correlation. Provider-specific score maps also transfer unevenly. By contrast, the fixed margin stays positive in every block and model after deduplication.

The secondary monotone projection reduces score MAE from 0.548 to 0.491 on the 32-paper development set and from 0.626 to 0.583 on the cross-year frozen test. Because the projection does not change ranking or HESR decisions, the reviewer-seat result is independent of this component.

7 DISCUSSION

HESR gives a conditional answer to whether AI can replace human peer review. One AI reviewer is not equally reliable on every paper, but its ordinal recommendation exposes a stable boundary region where replacement is least defensible. Routing that region to people yields approximately 55% pooled AI coverage and recommendation behavior comparable to another human draw on the selected papers.

The method's substantive insight is that a low-capacity assignment rule transfers better than a learned reviewer policy in this setting. The development study considered issue weights, specialist roles, fitted agreement probabilities, provider-specific mappings, and repeated calls. Their gains depended on the batch, scoring model, or added computation. The fixed margin instead survives changes in model family, year, conference outcome, and human-review consensus without refitting. Figure 2 also shows why HESR should be understood as boundary exclusion rather than unlimited trust in model confidence: rare extreme scores remain fallible.

HESR operates at the reviewer-seat level. Its target is the recommendation distribution that shapes conference decisions, while review-text factuality, rhetorical robustness, and area expertise remain separate quality dimensions. This separation makes the deployment rule concrete: one model can supply a review slot on selected papers, and human attention is concentrated where the model's recommendation is most likely to diverge from human judgment.

8 CONCLUSION

AI–human review disagreement is structured rather than uniform. A fixed half-level margin around the human recommendation boundary identifies model judgments that more closely match individual human reviewers. Across 99 papers, four model families, six evaluation blocks, and both decision directions, HESR raises recommendation agreement from 67.6% to 73.6% at 54.8% coverage while increasing historical decision accuracy. The result supports selective replacement by a single AI reviewer and a concrete allocation rule for preserving human review where model judgments are least stable.

AI USE STATEMENT

Generative AI is both the subject and an experimental instrument of this study. Qwen3.8-Max, GLM5.3, GPT-5.6-Sol, and DeepSeek DS4Pro produced the peer-review judgments evaluated in the paper. Generative AI tools also proposed and critiqued candidate hypotheses and experimental designs, implemented author-selected evaluation protocols, supported public-data processing and analysis, assisted result interpretation and translation, generated figures, and checked literature and citations. The authors formulated the research question, selected and froze the reported protocols, inspected source records and raw outputs, reproduced the numerical analyses, adjudicated the interpretation, and approved the final manuscript. The authors take responsibility for the complete submission.

REPRODUCIBILITY STATEMENT

The accompanying reproducibility package contains the review prompt, output schema, model-identity records, frozen protocols, executable analysis and figure code, and de-identified paper-level derived records for the 99-paper pooled analysis and 40-paper frozen tests. It excludes paper text, human-review prose, raw model-review text, titles, author identities, and original OpenReview identifiers.

REFERENCES

Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the Machine Learning Review Process Become More Arbitrary as the Field Has Grown? The NeurIPS 2021 Consistency Experiment. arXiv preprint, 2023. URL https://arxiv.org/abs/2306.03262.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-Agent Review Generation for Scientific Papers. arXiv preprint, 2024. URL https://arxiv.org/abs/2401.04259.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing. EMNLP, 2024. URL https://aclanthology.org/2024.emnlp-main.292/.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. NLPeer: A Unified Resource for the Computational Study of Peer Review. ACL, 2023. URL https://aclanthology.org/2023.acl-long.277/.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing Review Generation Through Prompt Generation. arXiv preprint, 2024. URL https://arxiv.org/abs/2402.10886.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. ICML, PMLR 97, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. ICML, PMLR 70, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

JerMa88. ICLR Peer Reviews 2024. Hugging Face dataset mirror, 2024. URL https://huggingface.co/datasets/JerMa88/ICLR_Peer_Reviews_2024.

JerMa88. ICLR Peer Reviews 2025. Hugging Face dataset mirror, 2025. URL https://huggingface.co/datasets/JerMa88/ICLR_Peer_Reviews_2025.

JerMa88. ICLR Peer Reviews 2026. Hugging Face dataset mirror, 2026. URL https://huggingface.co/datasets/JerMa88/ICLR_Peer_Reviews_2026.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement. ICLR, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/08dabd5345b37fffcbe335bd578b15a0-Abstract-Conference.html.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. NAACL-HLT, 2018. URL https://aclanthology.org/N18-1149/.

Lingyao Li, Junjie Xiong, Changjia Zhu, Runlong Yu, Chen Chen, Junyu Wang, Renkai Ma, and Zhicong Lu. LLM-as-a-Reviewer: Benchmarking Their Ability, Divergence, and Prompt Injection Resistance as Paper Reviewers. arXiv preprint, 2026a. URL https://arxiv.org/abs/2605.25415.

Ming Li, Chenguang Wang, Xirui Li, Xinyue Zeng, Dianqi Li, Peng Shi, Dawei Zhou, and Tianyi Zhou. How Can Rhetoric Reward-Hack AI Reviewers? Dissecting Rhetorical Sensitivity in AI-Based Peer Review. arXiv preprint, 2026b. URL https://arxiv.org/abs/2608.08975.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv preprint, 2023. URL https://arxiv.org/abs/2310.01783.

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. Bridging Human and LLM Judgments: Understanding and Narrowing the Gap. Advances in Neural Information Processing Systems 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/160f484617c9e3ce84332d8cbc8ba62b-Abstract-Conference.html.

Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. ReviewerToo: Should AI Join The Program Committee? A Look At The Future of Peer Review. arXiv preprint, 2025. URL https://arxiv.org/abs/2510.08867.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. CycleResearcher: Improving Automated Research via Automated Review. ICLR, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0a48036026dc7946ef6033ae14719cc5-Abstract-Conference.html.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. Benchmarking LLMs' Judgments with No Gold Standard. ICLR, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a9b0e4e205bdf232da9f74bfb9469539-Abstract-Conference.html.

A REPRODUCIBILITY

The reproducibility package provides the review prompt, output schema, model-identity records, frozen protocols, analysis code, sanitation and input-audit tools, figure scripts, and de-identified derived records. Its validation procedures check identifiers, score ranges, probability dimensions, probability sums, and packet hashes; a single command reproduces the central estimates reported for all four evidence sets. Each model uses a separate output path, and each paper is reviewed in a fresh context.

The 454-paper corpus registry deduplicates OpenReview identifiers across executed analyses, model outputs, training pairs, and confirmation sets. The HESR pooled analysis separately deduplicates its six four-model blocks and retains exactly four model judgments per included paper.

The cross-year frozen test fixes the model identities, 24-paper manifest, input hashes, HESR threshold, four success gates, and 10,000-replicate bootstrap before outputs are inspected. Its sample balances year, outcome, and human-review consensus. The human-reference robustness addendum is timestamped before any model output and does not change the primary sample or method.

B SCORE HARMONIZATION

Recommendation labels from recent ICLR forms are mapped as $\{1, 3, 5, 6, 8, 10\} \mapsto \{0, 1, 2, 3, 4, 5\}$. The corresponding anchors are strong reject, reject, marginally below, marginally above, accept, and strong accept; level 3 is therefore the common accept/reject boundary. Models receive these six anchors and may return a continuous score on $[0, 5]$. HESR uses that original score. The optional projection $0.7102497029r + 0.9020741672$ is pooled across models, has positive slope, and changes neither paper order nor the selection rule.

C DEVELOPMENT CONTROLS

The method-development program tested criterion-weight transfer, review-text adapters, paper-text priors, issue-level vetoes, specialized review roles, same-model repetition, probability calibration, and model-specific score mapping. The main paper reports controls that identify the effective component. The accompanying methods package provides the configurations and analysis procedures used to compare these designs.

D ADDITIONAL VERIFICATION ANALYSES

We ran four analyses that address the estimand and evidence questions directly. First, the two frozen tests contain 40 papers and 160 model–paper judgments. HESR retained 98 judgments (61.25% coverage) and increased AI–human agreement from 67.79% to 72.70%, a gain of 4.91 percentage points (95% paper-cluster interval, $[-0.49, 10.90]$). A selector retaining the same counts within model and recommendation direction reached 67.77%, leaving a 4.93-point margin-specific difference.

Second, on the retained frozen-test assignments, AI–human agreement was 72.70% and same-paper human–human agreement was 65.61%. Their difference was 7.09 points. Under the prespecified 5-point non-inferiority tolerance, the one-sided 95% paper-cluster bootstrap lower bound was -2.27 points; 97.97% of 10,000 bootstrap draws exceeded the non-inferiority boundary of -5 points.

Third, we recomputed the frozen analysis with equal weight per paper after averaging the selected model judgments within each paper. Thirty-eight of 40 papers retained at least one judgment. Agreement was 71.80% for HESR and 69.39% for all four model judgments on the same papers, a 2.41-point difference (95% paper-cluster interval, $[-3.69, 8.42]$). The point estimate therefore remains positive when papers, rather than retained reviewer seats, receive equal weight.

Finally, the cross-year frozen test produced positive seat-weighted gains in both the 2024 and 2025 strata (5.48 and 2.38 points). Before evaluation, every model packet passed the sanitation pipeline. The 40 frozen-test packets also passed an independent audit for publication, decision, review, identity, and stable-identifier cues. Transformation manifests record input and output hashes; eight generic language matches in the frozen set were verified as scientific body text.