From AI Reviewers to Evidence Assistants: Quantifying the Human-AI Responsibility Boundary in Peer Review

Zhouyang Wang[1,3*], Qiujie Xie[2,1*], Minjun Zhu[2,1*†], Shichen Li[1,3], Shulin Huang[1], Han Cui[1], Yiran Ding[1], Panzhong Lu[1], Zhenhao Liu[1], Fuchen Shen[1], Junshu Pan[1], Dalv Yin[1], Ke Sun[1], Zhiyuan Ning[1], Yixuan Weng[1], Peifeng Li[3], Yue Zhang[1]✉

[1]Engineering School, Westlake University, [2]Zhejiang University, [3]Soochow University

***: Equal contribution †: Project leader. Correspondence author✉:** zhangyue@westlake.edu.cn

Abstract

The rapid growth of AI conference submissions is putting new pressure on peer review. AI reviewer systems are increasingly proposed as support, but prior work leaves unresolved what responsibility their outputs should carry when they can surface useful critiques yet remain risky as independent judgments. We frame this as a responsibility-boundary problem. Using 600 ICLR 2026 submissions, 2231 human review traces, and 3,600 AI reviews, we operationalize this boundary through usable feedback, score use, panel breadth, and grounded synthesis. The results show that AI can prepare candidate critiques, organize evidence, and improve feedback, while scoring, independent panel judgment, high-level synthesis, and final responsibility should remain human-led. Motivated by this boundary, we develop **Review Copilot**, a workflow in which AI suggestions are inspected, edited, or rejected by human reviewers and provide neither official scores nor recommendations. In an initial controlled reviewer-in-the-loop study, Human+AI reviews improve actionability, evidence support, and professionalism relative to standalone baselines while preserving human authorship of scores and recommendations. Our results point toward a review paradigm in which AI expands the space of evidence-grounded critique, while humans remain responsible for judgment, synthesis, and accountability.

1 Introduction

The bottleneck in scientific publication is shifting from writing papers to reviewing them. AI tools for literature search, experimentation, and even automatic scientific discovery (Xie et al., 2025; Hao et al., 2026; Lu et al., 2024; Weng et al., 2026b) are making it quicker to produce technically plausible submissions (statistic in Figure 1). But **the cost of evaluating these submissions has not fallen accordingly.** High-quality peer review still requires scarce expert attention (Singh, 2025; Wei et al., 2025). Human reviewers must judge the correctness, novelty, significance, reproducibility, and venue fit under tight deadlines.

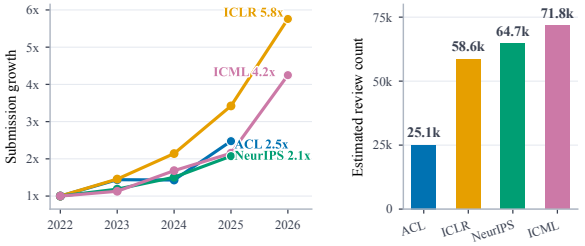

Figure 1: Major venues have expanded reviewer pools to handle rising submissions, yet this expansion remains constrained by the availability of qualified experts.

This pressure has accelerated research on **AI-assisted peer review**, utilizing LLMs as review generators, critique providers and evidence grounders (Zhuang et al., 2025; Mann et al., 2026; Zhu et al., 2025b). Yet numerous studies also make clear that **AI review cannot simply replace human judgment.** LLM-generated reviews can be brittle, biased, insufficiently grounded, and misaligned with community norms (Shin et al., 2025; Ye et al., 2024; Zhu et al., 2025a; Kim et al., 2026; Baumann et al., 2026; You et al., 2026). Large-scale monitoring further suggests that AI-modified review is already entering conference review, while policy analyses and prompt-injection studies show that unclear provenance and responsibility can create concrete governance risks (Collu et al., 2026; Theocharopoulos et al., 2025).

Major venues have therefore begun to explore bounded forms of machine assistance rather than fully automated reviewing. ICLR 2025 evaluated LLM feedback to reviewers while ICML 2026 offered the Paper Assistant Tool as private pre-submission feedback for authors (Jayaram et al., 2026). These efforts point toward a new **human-AI review paradigm** in which the key question is not whether AI should enter peer review, but **which responsibilities AI can support and which judgments must remain human-led.**

To explore this question, we first construct a large-scale comparison of human and AI reviews. Our evaluation dataset (Section 3) contains 600 ICLR 2026 submissions (OpenReview, 2026), sampled to balance final decisions while preserving variation across review ratings, together with 2,231 human reviews and 3,600 AI reviews. The AI reviews are generated by crossing two frontier foundation models (OpenAI, 2026; DeepSeek-AI, 2026) with three representative agentic review workflows (Weng et al., 2026a; Li et al., 2026; Jin et al., 2024), yielding six AI reviewer configurations.

We therefore evaluate AI review through four responsibilities (Section 4): (1) whether AI can produce professional and usable review artifacts; (2) whether its ratings can serve as calibrated reviewer scores; (3) whether multiple AI reviews can count as independent panel voices; and (4) whether grounded AI critiques carry the synthesis required for judgment. Across these evaluations, the lesson is not that AI lacks useful review content, but that useful content should not be mistaken for reviewer authority. Current systems can produce inspectable critique candidates, yet their scores are compressed and system-dependent, their reports add limited independent panel breadth, and their grounded critiques still require human synthesis. The resulting boundary is therefore artifact-level assistance under human control, not delegation to AI as a calibrated scorer, or an autonomous judge.

Motivated by this boundary, we build Review Copilot (Section 5) as a human-led workspace for evidence-centered reviewing. The prototype supports the parts of reviewing that precede judgment: helping reviewers interrogate the paper through grounded questions, recover supporting passages, surface relevant context, and organize candidate strengths, weaknesses, and questions into an inspectable critique space. Every AI-generated item remains provisional until the reviewer inspects, revises, or rejects it.

Our contributions are threefold: **(1)** using 600 ICLR 2026 submissions, 2,231 official human reviews, and 3,600 AI reviews, we **separate critique quality from review responsibility** and map where current AI reviewers can support peer review without being granted reviewer authority. **(2)** Our experiments **define a bounded role for AI in peer review**: it can support critique generation and evidence checking, but should not assume scoring, panel judgment, or final accountability. **(3)** We **introduce Review Copilot as an initial prototype** of the resulting human–AI review paradigm, where AI-generated critiques are inspected, revised, and selectively integrated by human reviewers. In our reviewer-in-the-loop study, the resulting reviews outperform both standalone human reviews and standalone AI reviews on the evaluated review-quality dimensions.

2 Related Work

AI Reviewer Systems Early LLM-based review systems treated reviewing largely as text generation, producing full reviews or author feedback (Du et al., 2024; Zhu et al., 2025b; Jin et al., 2024; Chang et al., 2025; Zeng et al., 2025). Recent work instead structures reviewing through rubric guidance, evidence grounding, tool checks, role specialization, and multi-agent interaction (Li et al., 2026; Gao et al., 2025; Kumar et al., 2026; Goyal et al., 2026; Taechoyotin and Acuna, 2026; Idahl and Ahmadi, 2025). This progression has made **AI review increasingly integrated into human review workflows**. Large-scale monitoring has detected AI-modified language in conference reviews (Liang et al., 2024), while major venues have begun testing bounded forms of assistance (Thakkar et al., 2025; Biswas et al., 2026; Jayaram et al., 2026). These developments show that AI review is moving from isolated generation systems toward real review infrastructures, **making it necessary to examine how their outputs behave inside human review processes.**

AI Review in Practice Despite substantial progress in AI reviewer systems, existing evidence cautions against treating LLM reviews as substitutes for reviewer judgment. LLM reviews may miss novelty, miscalibrate scores, reproduce biases, converge across generated reports, and remain vulnerable to hidden instructions or prompt-injection attacks (Renata and Lee, 2025; Kim et al., 2026; Shin et al., 2025; Ye et al., 2024; Zhu et al., 2025a; Pataranutaporn et al., 2025; Baumann et al., 2026; Theocharopoulos et al., 2025). Recent work therefore emphasizes measurement, disclosure, accountability, verification, and process design rather than autonomous delegation (Zhuang et al., 2025; Wu et al., 2026; Mann et al., 2026; Yun et al., 2026; Zhang and Abernethy, 2026; You et al., 2026).

However, these work leave underexplored the central question in the emerging human–AI review paradigm: **which reviewing responsibilities AI**

can support and which judgments must remain human-led. In this paper, we address this gap from a responsibility-boundary perspective through a large-scale analysis of 2,231 human reviews and 3,600 AI reviews, together with a reviewer-in-the-loop study of a Review Copilot prototype.

3 Experimental Settings

We construct the evaluation dataset from the full ICLR 2026 paper set crawled from OpenReview, including official reviews, reviewer ratings, and final decisions. We first partition papers by final decision into Oral, Poster, and Reject groups. Within each decision group, we divide papers into 0.5-point bins according to their mean official review rating and sample from these bins to obtain 200 papers per decision group. This design keeps the corpus balanced across outcomes while retaining papers across the rating spectrum, including both clear and borderline cases.

To generate the AI reviews, we treat each reviewer as a pairing of a foundation model and a review workflow, allowing us to separate model-level effects from workflow-level effects. For each paper, we generate six AI reviews by pairing each foundation model with each workflow. The foundation models are GPT-5.5 (OpenAI, 2026) and DeepSeek-V4-Pro (DeepSeek-AI, 2026); the review workflows are DeepReviewer-2.0 (Weng et al., 2026a), ReviewGrounder (Li et al., 2026), and AgentReview (Jin et al., 2024). We normalize all generated reports into a shared review schema consisting of Summary, Strengths, Weaknesses, Questions, Rating, and Decision. The final corpus contains 2,231 human reviews and 3,600 AI reviews.

4 Responsibility Boundaries of AI in human-AI review paradigm

4.1 Review Usability

Research Question. In this experiment, we evaluate whether AI can generate usable review artifacts rather than fluent text alone. We define usability by two requirements: a review should recover observable decision signals and should provide feedback that is actionable, grounded, and professional in tone. We therefore ask: *do human and AI reviews differ in how well they recover observable decision signals and meet basic professional standards?*

Evaluation Metrics. We compute decision alignment at the review level. Final decisions are binarized as accept for oral and poster papers and reject for rejected papers. For human reviews, we derive a reviewer-level decision from the official rating, treating ratings of 6 or higher as accept recommendations and ratings below 6 as reject recommendations. AI decisions are taken from the normalized Decision field. Acc and accept-class F1 are then computed against the binary final outcome.

The diagnostics target different parts of the review. *Actionable* is judged at the question-unit level and asks whether the feedback contains both a concern and a rebuttal-addressable suggestion. *Unsupported* is judged over weakness units after retrieving the top-3 paper evidence chunks, and counts unsupported or contradicted factual claims among all units. *Professionalism* is judged at the whole-review level over Summary, Strengths, Weaknesses, and Questions, and marks reviews as unprofessional if they contain major civility or abuse issues.

Experimental Results. As shown in Table 1, AI reviews are professionally usable. Several AI configurations recover observable outcome signals: GPT-5.5 with DeepReviewer2 obtains the highest accuracy, while DeepSeek with DeepReviewer2 obtains the highest accept-class F1. At the same time, most AI reviews satisfy basic professional standards under the judge diagnostics. They maintain near-perfect professionalism, produce lower unsupported-claim rates than human reviews, and often provide more rebuttal-addressable questions than the human average.

These results indicate that AI reviews can produce professionally usable artifacts that contain observable decision signals and meet basic standards of actionability, factual support, and tone. However, **this competence is unevenly distributed** across systems. DeepReviewer2 is strongest in decision alignment but weaker in actionability, whereas ReviewGrounder and AgentReview are often more actionable but less reliable as decision predictors. Therefore, the result establishes necessary point. Current AI systems can produce usable review artifacts, but artifact usability alone does not confer reviewer authority. The remaining analyses ask whether this usability extends to the responsibilities that make review judgment difficult to delegate.

4.2 Rating Calibration

Research Question. The previous experiment shows that AI systems can produce professionally usable review artifacts, but artifact-level competence does not imply reviewer authority. In peer

Model	System	Acc	F1	Actionable ↑	Unsupported ↓	Professionalism ↑
GPT-5.5	DeepReviewer2	**0.870**	0.901	0.386	**0.095**	0.998
GPT-5.5	ReviewGrounder	0.805	0.863	0.791	0.110	**1.000**
GPT-5.5	AgentReview	0.538	0.638	0.576	0.225	0.992
DeepSeek	DeepReviewer2	0.769	**0.905**	0.420	0.246	0.922
DeepSeek	ReviewGrounder	0.718	0.723	0.848	0.143	**1.000**
DeepSeek	AgentReview	0.530	0.623	**0.875**	0.215	0.931
Human	–	0.726	0.711	0.770	0.379	0.938

Table 1: **Evaluation of human reviews and AI reviews generated by different combinations.** Acc and F1 compare review decisions with final paper outcomes. We compute Actionable, Unsupported and Professionalism using an LLM-as-judge paradigm, with the prompt provided in the Appendix C.

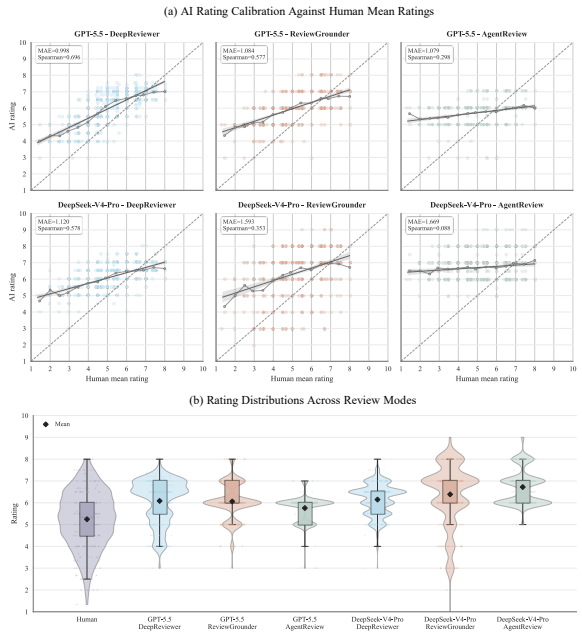

Figure 2: Calibration of AI ratings versus human expert judgments.

review, a rating is a process signal used by area chairs to compare reviewer stances, identify borderline papers, and decide where discussion should focus. Therefore, if AI is to take on any responsibility beyond drafting or organizing feedback, its scores must be calibrated to human reviewer judgments. We therefore ask: *do AI reviews use the rating scale in a way that is calibrated to human reviewer judgments?*

Evaluation Metrics. We use the mean human rating for each paper as the paper-level reference score produced by the review process. Each AI rating is compared with this reference score using MAE and Spearman correlation, while the full score distributions are used to examine how each AI system occupies the rating scale. The calibration view in Figure 2 (a) tests whether AI scores track human score differences across papers; the distributional view in Figure 2 (b) examines whether their use of the rating scale matches the range and concentration of human ratings.

Experimental Results. Figure 2 shows that AI ratings are only partially calibrated to human reviewer judgments. In the calibration plots, even the strongest setting, GPT-5.5 with DeepReviewer2, reaches only a moderate Spearman correlation with human mean ratings ($\rho = 0.696$), and the remaining systems show weaker rank alignment. More importantly, the fitted curves are much flatter than the diagonal reference line. When human reviewers assign high mean ratings, AI systems do not increase their ratings proportionally. This suggests that AI systems have limited sensitivity to the absolute meaning of the rating scale: they distinguish papers only weakly and tend to avoid both strongly negative and strongly positive judgments.

The distributional view in Figure 2 (b) confirms this pattern. Human ratings occupy a broad range of the scale, with visible mass from low to high scores. By contrast, most AI ratings concentrate around middle or moderately positive values, typically between 5 and 7, with much thinner use of the lower end of the scale.

These results suggest a clear responsibility boundary. Although the previous experiment showed that AI systems can generate professionally usable review artifacts, their ratings do not reliably encode the same process signal as human reviewer ratings. Therefore, AI should not be assigned score-level authority in human–AI reviewing. Final recommendations, calibration across papers, and prioritization of borderline cases should remain under human responsibility.

4.3 Perspective Breadth

Research Question. Although AI ratings are not sufficiently calibrated to support score-level authority, AI may still be useful to broaden the technical perspectives during review. In peer review, a panel

is valuable because different reviewers attend to different assumptions, missing baselines, data risks, methodological weaknesses, and failure modes before the area chair adjudicates them. We therefore ask: *do multiple AI-generated reviews provide distinct panel-level perspectives, or do they largely repeat the same evaluative concerns?*

Evaluation Metrics. We measure whether different reviews of the same paper provide diverse perspectives by comparing their semantic similarity. For each paper p, review section k, and pair of reviews i, j, we encode the corresponding section text $r^{(p,k)}$ using Qwen3-8B-Embedding f and compute cosine similarity:

$$s_{ij}^{(p,k)} = \frac{f(r_i^{(p,k)}) \cdot f(r_j^{(p,k)})}{\|f(r_i^{(p,k)})\| \|f(r_j^{(p,k)})\|}, \quad (1)$$

where **a higher $s_{ij}^{(p,k)}$ indicates greater semantic overlap between two reviews.** For example, they tend to identify similar strengths, weaknesses, or technical concerns.

We then compare similarity within human review panels and within AI-generated review panels. *Human Sim.* is the average similarity between pairs of human reviews written for the same paper. *AI Sim.* averages it over the 15 pair types induced by the six AI configurations. The standard deviation after $\pm$ is computed across these 15 pair-type averages, indicating how much similarity varies across AI-system pairs. We further report

$$\Delta = \text{AI Sim.} - \text{Human Sim.}, \quad (2)$$

where a positive Δ means that AI reviews are more semantically similar to each other than human reviews are.

To diagnose the source of AI-review convergence, we compare two types of AI-review pairs. Let S_{F} denote the mean similarity of pairs that share the same review workflow but use different base models, and let S_{M} denote the mean similarity of pairs that share the same base model but use different workflows. We define

$$\Delta\mathrm{F{-}M} = S_{\mathrm{F}} - S_{\mathrm{M}}. \quad (3)$$

A positive Δ F–M means that shared review workflows contribute more to convergence than shared base models.

Experimental Results. Table 2 shows that AI-generated reviews are consistently more semantically overlapping than human reviews across all review sections. The gap is small for Summary, where both humans and AI systems are expected to describe the same paper content. However, the gap becomes much larger in the evaluative sections. For Strengths and Weaknesses, AI reviews show substantially higher similarity than human reviews, with $\Delta = 0.187$ and $\Delta = 0.203$, respectively. This suggests that AI systems not only summarize papers in similar ways, but also tend to identify similar strengths, weaknesses, and rebuttal-facing concerns. The positive Δ F–M values further point to review workflows as a major source of AI-review convergence, rather than base models alone; full pair-type statistics are reported in Appendix D.

Section	Human	AI	Δ	Δ F–M
Summary	0.815	0.835 ± 0.037	0.020	0.066
Strengths	0.648	0.835 ± 0.033	0.187	0.016
Weaknesses	0.631	0.835 ± 0.041	0.203	0.021
Questions	0.503	0.776 ± 0.050	0.274	0.065

Table 2: Semantic overlap among human reviews and AI reviews. Δ is AI similarity minus human similarity, and Δ F–M measures whether convergence is driven more by shared frameworks than by shared base models.

This result does not imply that human disagreement is always valuable. Some disagreement may reflect noise, or reviewer-specific bias. The point is narrower: **peer review needs room for multiple plausible technical readings before an area chair adjudicates them.** On this criterion, multiple AI reviews do not provide the same kind of panel breadth as multiple human reviews. They may help populate a review workspace with candidate concerns, but the high similarity across AI reviews means that they should not be counted as independent panel voices. In the human–AI review paradigm, perspective breadth therefore remains a human-led responsibility.

4.4 Grounded Synthesis

Research Question. The preceding experiments examine AI reviews through external signals: review usability, rating calibration, and cross-review diversity. We now look inside the review text. A review can be unfolded as a sequence of evaluative units that summarize evidence from the submission and gradually form an implicit judgment. This view motivates our focus on **grounded synthesis**, which combines two separable requirements: *local evidence grounding*, i.e., whether a review unit is anchored in concrete evidence from the target submission rather than generic paper-like content, and

meta-level evaluative synthesis, i.e., whether the review moves beyond restating local evidence to assess what that evidence implies for the paper's claims, weaknesses, and overall merit. We therefore ask: *how do human and AI review trajectories differ in balancing paper-specific grounding with evaluative synthesis beyond local evidence?*

Evaluation Metrics. For each paper p and review group g, we split the critique into an ordered set of units $R_{p,g} = \{r_1, r_2, \ldots, r_n\}$. Each unit is a sentence, bullet, or short paragraph from the critique-bearing sections. We embed each unit with Qwen3-8B-Embedding to obtain a semantic trajectory $X_{p,g} = [x_1, x_2, \ldots, x_n]$, where $x_i = f(r_i) \in \mathbb{R}^d$. This trajectory is the external semantic trace of how the review text organizes evaluative information under the constraint of paper evidence.

We then characterize grounded synthesis along two axes. The first axis measures **local evidence grounding**: whether a review unit is more semantically supported by the target paper than by similar non-target papers. Let $\mathcal{T}_{p,i}^{(k)} = \mathrm{TopK}_k(P_p; r_i)$ be the top-k target-paper chunks for unit r_i. We compute target-paper support as:

$$T_i = \frac{1}{k} \sum_{c \in \mathcal{T}_{p,i}^{(k)}} \cos\big(f(r_i), f(c)\big). \tag{4}$$

We compare this with balanced hard-negative pools sampled from semantically similar non-target papers. Let $\mathcal{H}_{p,b,i}^{(k)} = \mathrm{TopK}_k(H_{p,b}^{\mathrm{bal}}; r_i)$; using $k = 3$ and $B = 20$, the negative support is:

$$N_i = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{k} \sum_{c \in \mathcal{H}_{p,b,i}^{(k)}} \cos\big(f(r_i), f(c)\big). \tag{5}$$

The unit-level support margin is $M_i = T_i - N_i$. For the paper–group level used in the regime map, we average unit-level margins: $M(p,g) = \frac{1}{|R_{p,g}|} \sum_{r_i \in R_{p,g}} M_i$. A higher value indicates that the critique is more specifically anchored in the submitted paper rather than in generally similar papers.

The second axis measures **meta-level evaluative synthesis beyond local evidence**. For each unit, we retrieve the nearest target-paper chunk $c_i^+ = \arg\max_{c \in P_p} \cos(f(r_i), f(c))$, and define the evidence-conditioned residual as $u_i = f(r_i) - f(c_i^+)$. Based on the residual trajectory $U_{p,g} = \{u_i\}_{i=1}^n$, we compute residual $V/unit$:

$$V_{p,g}^{\mathrm{res}} = \frac{1}{2n} \log\det\big(I + \alpha \widetilde{U}_{p,g} \widetilde{U}_{p,g}^{\top}\big), \tag{6}$$

where $\widetilde{U}_{p,g}$ is the centered residual matrix. This residual axis captures the transformation between evidence localization and final judgment.

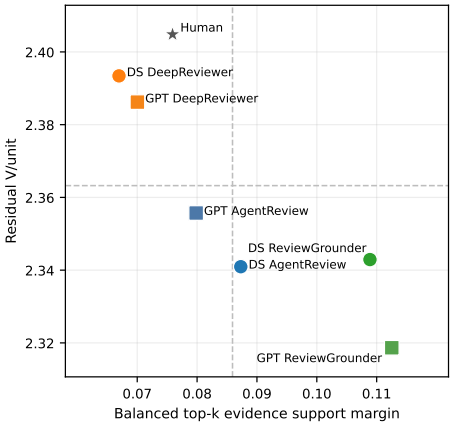

Figure 3: Grounded-synthesis regime map under the full-critique condition. The x-axis measures paper-specific semantic support, and the y-axis measures post-local-evidence residual dispersion. The map is diagnostic rather than a quality ranking.

Experimental Results. Figure 3 separates review groups by how they balance paper-specific grounding and evaluative synthesis. ReviewGrounder lies farthest along the evidence-support axis, consistent with its grounding-oriented design, but its lower residual movement suggests a more constrained semantic trajectory. Human reviews occupy the complementary regime: they retain the highest residual variation while showing only moderate paper-specific support. This should not be read as weaker human critique; rather, human reviewers often compress the evidence-to-judgment chain, stating evaluative conclusions without making every supporting manuscript passage explicit. DeepReviewer2 retains more residual movement but is less locally grounded, while AgentReview falls closer to the middle.

These results suggest that AI is better suited to anchoring and inspecting candidate critiques against paper evidence than to taking over meta-level reviewer judgment. The responsibility boundary is therefore more clear: AI can support evidence-grounded critique, but humans should remain responsible for deciding which evidence matters and how it should affect the final recommendation. More results are reported in Appendix E.

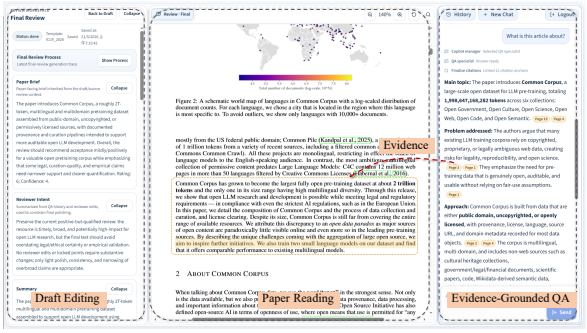

Figure 4: Review Copilot workspace used in the controlled review study. The interface integrates paper reading, evidence-grounded QA, draft editing, evidence inspection, allowing reviewers to inspect AI suggestions before incorporating them into the final report.

5 Exploring the New Paradigm of Human-AI Review

Building on Section 4, we operationalize this boundary in a controlled Review Copilot study, where AI supports paper-grounded fact collection, relevant-context organization, and candidate critique structuring under human control rather than generating a rating or decision.

As shown in Figure 4, Review Copilot implements an evidence-grounded assistance workflow. Reviewers ask questions while reading the paper, and the system returns paper-grounded answers with supporting evidence. Reviewers then inspect these answers, revise candidate critique points, and decide which content enters the final review. The system provides no official score or recommendation, so the final review, score, and recommendation remain human-authored. The study evaluates whether this workflow improves the final review artifact while preserving human-led scoring, active reviewer mediation, and evidence-supported synthesis.

5.1 Experimental Setup

We invited 10 anonymized, experienced AI-conference reviewers. We sampled 40 papers from the 600-paper ICLR 2026 evaluation set, balanced between accepted and rejected papers. Each reviewer completed four reviews in Review Copilot, yielding 40 Human+AI reviews. All reviewers used the same assisted workflow, with protocol details reported in Appendix F, allowing us to compare their final reviews with the human and AI-only baselines. To keep this comparison model-controlled, Review Copilot uses GPT-5.5 as its underlying model, and the AI-only baselines in this section are averaged over the three GPT-5.5 workflow conditions.

Group	Acc. ↑	F1 ↑	Act. ↑	Unsup. ↓	Prof. ↑
AI	0.742	0.739	0.634	0.097	0.995
Human	**0.831**	**0.831**	0.769	0.194	0.956
Human+AI	0.800	0.800	**0.952**	**0.019**	**1.000**

Table 3: Human-AI cooperation improves review artifact quality in actionability and professionalism.

5.2 Professional Gains Under Human Control

We first test whether AI-generated material can improve review writing without transferring review responsibility to AI. In Review Copilot, reviewers use AI answers and candidate critique points only as intermediate material; the final review is written, revised, and submitted by the human reviewer.

Table 3 shows a clear artifact-level gain. Human+AI reviews are more actionable than both human-only and AI-only reviews, with an actionability score of 0.9516. They also contain far less unsupported content than human-only reviews (0.0185 vs. 0.1942) and reach the highest professionalism score. Meanwhile, their decision agreement remains close to the human baseline rather than following an autonomous AI decision pattern.

5.3 Score Responsibility Remains Human-Led

Section 4.2 showed that AI-only ratings are poorly calibrated: they shift upward and use a compressed score range. Review Copilot therefore does not provide an official score or recommendation; reviewers enter the score themselves. Table 4 shows that Human+AI scores do not follow the AI-only rating pattern. Human+AI scores have a much smaller mean shift (+0.14) and a less compressed scale (Scale Ratio = 0.74) than AI-only scores. Their MAE is also slightly lower than both AI-only and human-only scores. These results support the intended responsibility allocation. AI can assist the construction of written feedback, but the score should remain a human process signal. Review Copilot improves the review artifact without carrying AI-generated ratings into the decision pipeline.

5.4 Critique Formation Remains Human-Led

We next examine whether AI suggestions enter the final review unchanged or are actively mediated by reviewers. In Review Copilot, AI-generated content is only candidate material: reviewers may accept, revise, or reject each point before finalization. We use interaction logs to measure this pro-

Source	Score Source	MAE ↓	Mean Shift ↓	Scale Ratio
AI	AI	1.19	+0.85	0.44
Human	Human	1.15	0.00	1.00
Human+AI	Human	1.11	+0.14	0.74

Table 4: Human-led scoring avoids AI-only rating distortions. Shift: mean rating offset from the human baseline; Scale Ratio: ratio between each source's rating standard deviation and the human rating standard deviation.

Artifact	Accepted	Edited	Rejected	Med./Review
Summary	9	31	0	0.78
Strengths	64	84	9	**2.33**
Weaknesses	77	73	12	2.13
Questions	129	56	15	1.78

Table 5: Reviewer mediation of AI artifacts. Med./Review reports the average number of artifacts edited or rejected per final review.

cess. Table 5 shows that reviewers do not passively copy AI outputs. They revise most summaries and frequently edit or reject critique-bearing points, especially strengths and weaknesses. This mediation was not prompted as a requirement to add new critique. Reviewers were instructed to treat Review Copilot as optional assistance and could add, delete, rewrite, or lock points reflecting their own judgment (Appendix F). Thus, the observed edits and rejections indicate reviewer-authored critique formation rather than passive filtering of AI output.

5.5 Grounded Synthesis in Human-AI Review

Sections 5.2–5.4 show that Review Copilot improves the final review artifact while keeping responsibility with the reviewer. We now ask how this improvement is produced. The improvement cannot be simply attributed to AI injecting more information. As cautioned in Section 4.4, unconstrained residual semantic movement can reflect useful synthesis (e.g., integrating evidence) just as easily as unsupported drift or hallucination. In the Human+AI setting, the key question is whether human mediation concentrates this evaluative movement into directions robustly supported by the target paper.

To test this, we use two diagnostics built from the grounding and residual axes in Section 4.4. Support-Scaled Residual Volume (SSRV) measures how much evaluative movement remains after local evidence is accounted for, while weighting that movement by positive paper-specific support; it is high when a review develops judgments beyond restating evidence and those judgments remain anchored in the target paper. We report it together with grounded weighted effective dimension (GWD_{eff}), which measures how many supported semantic directions this evidence-weighted residual content occupies. A lower GWD_{eff} is not a standalone quality signal, but when paired with high SSRV, it indicates compact evidence-supported synthesis: the review concentrates substantial supported evaluative content into fewer, stronger critique directions. Full weighting, centering, and component diagnostics are given in Appendix E.3.

Group	SSRV ↑	GW D_{eff} ↓
Human	0.182	7.737
AI	0.201	8.149
Human+AI	**0.243**	**5.802**

Table 6: Support-scaled grounded synthesis. High SSRV indicates rich evidence-anchored evaluative volume, while a lower effective dimension (GWD_{eff}) reflects more compact semantic concentration.

As shown in Table 6, Human+AI reviews consistently achieve the highest support-scaled residual volume ($SSRV = 0.243$), outperforming both human reviews (0.182) and the AI mean (0.201). A more profound pattern emerges from the combination of this high volume with a lower supported effective dimension. Human+AI has lower grounded residual effective dimension ($GWD_{eff} = 5.802$) than both human review (7.737) and the AI mean (8.149), while having higher SSRV. We interpret this phenomenon as semantic distillation. The Human+AI review occupies fewer directions, but those directions carry more evidence-scaled evaluative content.

This pattern connects the semantic analysis to the artifact-level gains in Section 5.2. Review Copilot does not merely add more critique; it helps reviewers consolidate candidate evidence and critique points into a more compact, evidence-supported final review. Crucially, lower SSRV in human-only reviews does not imply inferiority. Human reviewers often rely on domain knowledge, implicit comparisons to prior work or common baselines, and expectations about evidence sufficiency that are not always stated as locally recoverable passages in the submitted paper. The result instead shows a shift in the review trajectory: Human+AI review moves part of the critique toward evidence-recoverable synthesis, while scoring and final responsibility remain human-led.

6 Conclusion

We studied AI-assisted peer review as a responsibility-boundary problem. Across human and AI reviews, we found that current review agents produced professional, actionable, and information-rich artifacts, but their scores remained system-dependent, their reports added limited independent panel breadth, and their grounded critiques still required human adjudication. These findings support a bounded role for AI in peer review: AI should assist reviewers by collecting paper-grounded evidence, organizing candidate concerns, and improving feedback under human control, rather than acting as an autonomous reviewer. Scoring, synthesis, panel judgment, and final responsibility should remain human-led.

Limitations

The main corpus comes from ICLR 2026, whose submission pool, rating conventions, review form, and decision process may differ from those of other venues. The LLM-as-judge audits depend on judge-model behavior and evidence retrieval quality, while embedding-based diversity and synthesis metrics cannot fully separate useful disagreement from noise or plausible critique from correct critique. We also evaluate only two foundation models, three agentic workflows, and a reviewer-in-the-loop study with 10 reviewers and 40 assisted reviews; broader deployment claims therefore require replication across venues, fields, reviewer populations, model families, interfaces, and policy settings.

Ethical Considerations

This work studies AI assistance in a high-stakes peer-review setting, and its results should not be used to justify autonomous AI reviewing. The central ethical risk is not the presence of AI assistance itself, but allowing generated artifacts to acquire reviewer authority without disclosure, verification, or accountability. Any deployment of systems like the ones studied here should preserve reviewer and area-chair responsibility, disclose AI assistance when required by venue policy, and prevent AI-generated scores or text from entering the decision process without human inspection. Our analyses use review-process data in aggregate rather than to evaluate individual authors or reviewers. In the reviewer-in-the-loop study, interaction logs are used to study workflow mediation, not reviewer ability. Future releases of data, prompts, logs, or model outputs should remove identifying information and respect venue policies on manuscript confidentiality, reviewer anonymity, and third-party model use.

References

Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors. 2024. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. 2026. Stop automating peer review without rigorous evaluation.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. 2026. Ai-assisted peer review at scale: The aaai-26 ai review pilot.

Houda Bouamor, Juan Pino, and Kalika Bali, editors. 2023. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore.

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.

Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors. 2025. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria.

Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors. 2025. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China.

Matteo Gioele Collu, Umberto Salviati, Roberto Confalonieri, Mauro Conti, and Giovanni Apruzzese. 2026. Misleading large language models used (or misused) in scientific peer-reviewing via hidden prompt-injection attacks. *ACM Trans. AI Secur. Priv.* Just Accepted.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. Technical report, DeepSeek-AI. Technical report.

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. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. Reviewagents: Bridging the gap between human and ai-generated paper reviews.

Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors. 2022. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. 2026. Scholarpeer: A context-aware multi-agent framework for automated peer review.

Qianyue Hao, Fengli Xu, Yong Li, and James Evans. 2026. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 649:1237 – 1243.

ICLR. 2022. Tenth annual international conference on learning representations (iclr) 2022 fact sheet.

ICLR. 2023. Eleventh annual international conference on learning representations (iclr) 2023 fact sheet.

ICLR. 2024. Twelfth annual international conference on learning representations (iclr) 2024 fact sheet.

ICLR. 2025. 13th annual international conference on learning representations (iclr) 2025 fact sheet.

ICLR. 2026. 14th annual international conference on learning representations (iclr) 2026 fact sheet.

ICML. 2024. 41st annual international conference on machine learning (ICML) 2024 fact sheet.

ICML. 2025. 42nd annual international conference on machine learning (ICML) 2025 fact sheet.

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.

Rajesh Jayaram, Vincent Cohen-Addad, Alekh Agarwal, Miroslav Dudik, Sharon Li, and Martin Jaggi. 2026. Icml experimental program using google's paper assistant tool (pat).

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

Seungone Kim, Dongkeun Yoon, Kiril Gashteovski, Juyoung Suk, Jinheon Baek, Pranjal Aggarwal, Ian Wu, Viktor Zaverkin, Spase Petkoski, Daniel R. Schrider, Ilija Dukovski, Francesco Santini, Biljana Mitreska, Yong Jeong, Kyeongha Kwon, Young Min Sim, Dragana Manasova, Arthur Porto, Biljana Mojsoska, Makoto Takamoto, Marko Shuntov, Ruoqi Liu, Hyunjoo Jenny Lee, Niyazi Ulas Dinç, Yehhyun Jo, Sunkyu Han, Chungwoo Lee, Huishan Li, Esther H. R. Tsai, Ergun Simsek, Khushboo Shafi, Yeonseung Chung, Jihye Park, Aleksandar Shulevski, Henrik Christiansen, Yoosang Son, Elly Knight, Amanda Montoya, Jeongyoun Ahn, Christian Langkammer, Heera Moon, Changwon Yoon, Nikola Stikov, Mooseok Jang, Edward Choi, Junhan Kim, Yeon Sik Jung, Woo Youn Kim, Jae Kyoung Kim, Ishraq Md Anjum, Hyun Uk Kim, Drew Bridges, Carolin Lawrence, Xiang Yue, Alice Oh, Akari Asai, Sean Welleck, and Graham Neubig. 2026. On the limits and opportunities of ai reviewers: Reviewing the reviews of nature-family papers with 45 expert scientists.

Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors. 2024. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand.

Komal Kumar, Aman Chadha, Salman Khan, Fahad Shahbaz Khan, and Hisham Cholakkal. 2026. Paper circle: An open-source multi-agent research discovery and analysis framework.

Zhuofeng Li, Yi Lu, Dongfu Jiang, Haoxiang Zhang, Yuyang Bai, Chuan Li, Yu Wang, Shuiwang Ji, Jianwen Xie, and Yu Zhang. 2026. Reviewgrounder: Improving review substantiveness with rubric-guided, tool-integrated agents.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *International Conference on Machine Learning*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *ArXiv*, abs/2408.06292.

Sebastian Porsdam Mann, Mateo Aboy, Joel Jiehao Seah, Zhicheng Lin, Xufei Luo, Daniel Rodger, Hazem Zohny, Timo Minssen, Julian Savulescu, and Brian D. Earp. 2026. Ai and the future of academic peer review.

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors. 2022. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland.

NeurIPS. 2022. 36th annual conference of neural information processing systems (neurips) 2022 fact sheet.

NeurIPS. 2023. 37th annual conference of neural information processing systems (neurips) 2023 fact sheet.

NeurIPS. 2024. 38th annual conference of neural information processing systems (neurips) 2024 fact sheet.

NeurIPS. 2025. 39th annual conference of neural information processing systems (neurips) 2025 fact sheet.

OpenAccept. 2026. ICML submission statistics.

OpenAI. 2026. Gpt-5.5 system card. System card, OpenAI.

OpenReview. 2026. ICLR 2026 conference submissions and proceedings. https://openreview.net/group?id=ICLR.cc/2026/Conference. Accessed: 2026-05-23.

Pat Pataranutaporn, Nattavudh Powdthavee, Chayapatr Achiwaranguprok, and Pattie Maes. 2025. Can ai solve the peer review crisis? a large scale cross model experiment of llms' performance and biases in evaluating over 1000 economics papers.

Vianney Renata and John Lee. 2025. Ai reviewers: Are human reviewers still necessary? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 69(1):338–342.

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors. 2023. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Mind the blind spots: A focus-level evaluation framework for LLM reviews. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35630–35656, Suzhou, China. Association for Computational Linguistics.

Bhavneet Singh. 2025. Epistemic destabilization: Ai-driven knowledge generation and the collapse of validation systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2387–2398.

Pawin Taechoyotin and Daniel E. Acuna. 2026. Remctx: Automated peer review via reinforcement learning with auxiliary context.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025.

Panagiotis Theocharopoulos, Ajinkya Kulkarni, and Mathew Magimai. Doss. 2025. Multilingual hidden prompt injection attacks on llm-based academic reviewing.

Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. 2025. The ai imperative: Scaling high-quality peer review in machine learning.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Zhiyuan Ning, Shichen Li, Panzhong Lu, Zhen Lin, Enhao Gu, Qiyao Sun, and Yue Zhang. 2026a. Deepreviewer 2.0: A traceable agentic system for auditable scientific peer review.

Yixuan Weng, Minjun Zhu, Qiujie Xie, QiYao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2026b. Deepscientist: Advancing frontier-pushing scientific findings progressively. In *The Fourteenth International Conference on Learning Representations*.

Sihong Wu, Owen Jiang, Yilun Zhao, Tiansheng Hu, Yiling Ma, Kaiyan Zhang, Manasi Patwardhan, and Arman Cohan. 2026. Can ai be a good peer reviewer? a survey of peer review process, evaluation, and the future.

Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, Jian Wu, and Yue Zhang. 2025. How far are ai scientists from changing the world?

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review.

Lei You, Lele Cao, and Iryna Gurevych. 2026. Preventing the collapse of peer review requires verification-first ai.

JungMin Yun, JuneHyoung Kwon, MiHyeon Kim, and YoungBin Kim. 2026. Position on llm-assisted peer review: Addressing reviewer gap through mentoring and feedback.

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.

Tianmai M. Zhang and Neil F. Abernethy. 2026. Reviewing scientific papers for critical problems with reasoning llms: Baseline approaches and automatic evaluation.

Changjia Zhu, Junjie Xiong, Renkai Ma, Zhicong Lu, Yao Liu, and Lingyao Li. 2025a. When your reviewer is an llm: Biases, divergence, and prompt injection risks in peer review.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025b. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332.

A License For Artifacts

The artifacts associated with this work include evaluation code, analysis scripts, evaluation prompts, derived statistics, generated AI-review outputs, and Review Copilot study materials. The evaluation prompts used for the LLM-as-judge audits are included in Appendix C; code and scripts will be released under the license specified in the public repository. OpenReview-derived materials, including manuscript PDFs, official reviews, ratings, and decisions, remain subject to OpenReview and venue terms. We therefore do not redistribute raw submissions, raw reviewer text, reviewer identities, or other sensitive metadata. Where policy permits, we release only paper identifiers, derived aggregate statistics, and scripts needed to reproduce the analyses from authorized data access.

Generated AI reviews and Review Copilot logs may contain manuscript-specific or reviewer-interaction information, so any release of these artifacts will be limited to de-identified, policy-compliant examples or aggregate statistics. Users of the artifacts are responsible for respecting venue policies on manuscript confidentiality, reviewer anonymity, third-party model use, and disclosure of AI assistance.

B Statistics on Review Loads

The average review load is uniformly calculated as

$$\text{AvgReviewLoad} = \frac{\text{Submissions} \times 3}{\text{Reviewers}}.$$

For ICLR 2026, the official average review load is used. For years in which the number of reviewers is not publicly available, the value is kept as "-". Table 7 reports the resulting conference-level statistics.

C LLM-as-Judge Prompts for Review Usability

This appendix reports the exact prompts used for the three LLM-as-judge diagnostics in Table 1. The prompts correspond to different review responsibilities and different input granularities. *Actionable* is evaluated on pre-segmented units from the Questions section. *Unsupported* is evaluated on fixed feedback units against retrieved paper evidence. *Professionalism* is evaluated once per full review using the Summary, Strengths, Weaknesses, and Questions sections.

C.1 Actionable Prompt

Figure 5 reports the exact prompt used for the Actionable metric.

C.2 Professionalism Prompt

Figure 6 reports the exact prompt used for the Professionalism metric.

C.3 Unsupported Prompt

Figure 7 reports the exact prompt used for the Unsupported metric.

D Review Similarity and Diversity Details

This appendix reports the detailed similarity statistics behind Table 2. We use the section-level cosine similarity $s_{ij}^{(p,k)}$ defined in Section 4.3 and aggregate it as follows. Human Sim. averages $s_{ij}^{(p,k)}$ over within-paper human-human reviewer pairs and then across papers. AI Sim. averages the 15 AI configuration-pair types formed by the six AI configurations.

Finally, we define Δ as AI Sim. $-$ Human Sim., and Δ F–M as the mean similarity of same-framework, cross-model pairs minus the mean similarity of same-model, cross-framework pairs.

Tables 8 and 9 use the same reporting schema. The unit column indicates whether the distribution is computed over human reviewer pairs or over AI configuration-pair means; n reports the number of such units. Configuration abbreviations use the model prefix G for GPT-5.5 and D for DeepSeek-V4-Pro, followed by ARev for AgentReview, DRev for DeepReviewer2, and RGrd for ReviewGrounder. For example, D-RGrd denotes DeepSeek-V4-Pro with ReviewGrounder.

E Details for Grounded Synthesis Metrics

This appendix provides the full metric definitions, implementation details, complete ranking tables, and additional diagnostic maps for grounded synthesis: how human and AI reviews balance local evidence grounding with meta-level evaluative synthesis beyond local evidence.

E.1 Balanced Top-k Evidence Support Margin

The Balanced Top-k Evidence Support Margin operationalizes the local-evidence grounding axis. It measures whether a review unit is more specifically supported by the target paper than by semantically similar evidence from non-target papers. For each

Actionable Prompt

```
You are an expert peer-review evaluation judge.

Evaluate actionable feedback only. In this metric, "actionable" means the feedback gives authors a clear rebuttal target
      by containing both:
(1) an explicit reviewer concern/rationale/question/evidence gap, and
(2) a concrete response path the authors can take during rebuttal.

The concern may be phrased as a question, a declarative critique, or a recommendation with an explicit rationale. It does
       not need to be a question sentence.

- The input contains a fixed list of pre-segmented feedback_units from the Questions section only.
- Do not split, merge, drop, reorder, or rewrite the input units.
- Return exactly one output item for each input unit, preserving feedback_unit_index, source_section, and split_method.
- Do not echo unit_text in the output.
- Mark is_actionable=true only if the unit itself contains BOTH:
  A. Concern/rationale: what the reviewer is uncertain about, skeptical of, unable to verify, finds unclear, believes is
      unsupported, sees as a confound, sees as an unfair comparison, or asks the authors to
explain.
  B. Concrete response path: what authors can do in rebuttal, such as clarify, justify, point to existing evidence,
      report already available results, explain a limitation, explain why a requested
comparison/experiment is not applicable, or address the concern with a concrete experiment, ablation, comparison, metric,
       analysis, or correction.
- A concern can be expressed declaratively, for example: "The current experiment leaves open whether X is caused by Y" or
       "The theorem assumes X, which may not match the experimental setting."
- Imperative wording alone does not make a unit false. If the unit states the concern/rationale motivating the requested
      action, mark true.
- Mark is_actionable=false if the unit only gives a requested edit, experiment, baseline, metric, or analysis without
      stating why it matters for a claim, evidence chain, comparison, assumption,
interpretation, or reviewer uncertainty.
- Mark is_actionable=false if the unit only states a concern but gives no concrete way for authors to respond.
- Do not infer an unstated concern from a useful, detailed, or specific revision suggestion.
- Do not judge factual correctness, paper grounding, or whether the suggestion is already addressed by the paper.
- Pure summary, pure praise, ratings, final decisions, broad topical headings, "No questions", "See weaknesses", and
      vague complaints should be marked is_actionable=false if they appear in the fixed input
list.

Examples:
- "Add stronger baselines." -> false, because it gives an action but no concern/rationale.
- "Add stronger baselines including S4, Mamba, and RetNet." -> false, because details alone do not state why the
      comparison matters.
- "Report sensitivity to hyperparameters." -> false, because it gives an action but no explicit concern.
- "The evidence is weak." -> false, because it states a concern but no concrete response path.
- "The theory is unclear." -> false, because it states a concern but no concrete response path.
- "Without stronger sequence-model baselines such as Mamba or S4, it is hard to tell whether the gains come from long-
      context capacity rather than the proposed mechanism; please add such comparisons or
justify why they are not applicable." -> true.
- "The current ablations do not isolate whether the retrieval module or reranking module drives the improvement; an
      ablation removing each component would clarify this." -> true.
- "The theorem assumes uniform context distributions, but the experiments use non-uniform latent POMDPs; please clarify
      whether the bound is expected to transfer or state the limitation." -> true.
- "Prediction improvements alone do not establish practical benefit for planning or control; include downstream planning
      evaluations or explain why open-loop prediction is sufficient." -> true.
- "Compare to Mamba because the architecture-superiority claim is unclear without sequence-model baselines." -> true.
- "Could the authors explain why S4/Mamba baselines are omitted?" -> true.
- "Discuss limitations more thoroughly." -> false.

Return JSON only, with this exact shape:
{
  "feedback_units": [
    {
      "feedback_unit_index": 1,
      "source_section": "questions",
      "split_method": "marker | paragraph | single_prose_block",
      "is_actionable": true,
      "reason": "..."
    }
  ]
}
```

Figure 5: Exact LLM-as-judge prompt for the Actionable metric.

Conference	Year	Submissions	Reviewers	AC / SAC	Avg. Load	Source
ACL	2022	3,378	2,323	363 / 82	4.36	Muresan et al. (2022)
	2023	4,864	4,490	438 / 70	3.25	Rogers et al. (2023)
	2024	4,835	4,209	718 / 72	3.45	Ku et al. (2024)
	2025	8,360	11,720	1,937 / 169	2.14	Che et al. (2025)
EMNLP	2022	4,190	3,828	297 / 70	3.28	Goldberg et al. (2022)
	2023	4,909	4,094	458 / 85	3.60	Bouamor et al. (2023)
	2024	6,395	10,309	1,458 / 99	1.86	Al-Onaizan et al. (2024)
	2025	8,174	13,048	1,989 / 168	1.88	Christodoulopoulos et al. (2025)
ICLR	2022	3,391	5,507	394 / 20	1.85	ICLR (2022)
	2023	4,938	5,734	413 / 40	2.58	ICLR (2023)
	2024	7,262	8,950	624 / 60	2.43	ICLR (2024)
	2025	11,603	18,325	823 / 71	2.60	ICLR (2025)
	2026	19,525	21,674	1,634 / 79	4.22	ICLR (2026)
NeurIPS	2022	10,411	10,406	742 / 82	3.00	NeurIPS (2022)
	2023	12,343	12,974	968 / 98	2.86	NeurIPS (2023)
	2024	15,671	13,640	1,393 / 195	3.45	NeurIPS (2024)
	2025	21,575	21,921	1,985 / 240	2.95	NeurIPS (2025)
ICML	2022	5,630	7,403	281 / 53	2.27	OpenAccept (2026)
	2023	6,358	6,053	504 / 47	3.15	OpenAccept (2026)
	2024	9,473	7,474	492 / 97	3.76	Fact sheet (ICML, 2024)
	2025	12,107	10,943	1,161 / 155	3.32	Fact sheet (ICML, 2025)
	2026	23,918	–	–	–	OpenAccept (2026)

Table 7: Submission and reviewing statistics of major conferences.

Section	Unit	n	Mean	Std.	Range
Summary	Human pairs	3442	0.815	0.085	0.421–0.974
Strengths	Human pairs	3442	0.648	0.109	0.274–0.926
Weaknesses	Human pairs	3442	0.631	0.102	0.223–0.863
Questions	Human pairs	3442	0.503	0.149	0.121–1.000

Table 8: Human-human similarity distribution by section. The distribution unit is an within-paper pair of human reviews.

Section	Unit	n	Mean	Std.	Range
Summary	Config. pairs	15	0.835	0.037	0.780–0.905
Strengths	Config. pairs	15	0.835	0.033	0.766–0.889
Weaknesses	Config. pairs	15	0.835	0.041	0.749–0.912
Questions	Config. pairs	15	0.776	0.050	0.673–0.862

Table 9: AI-AI similarity distribution by section. The distribution unit is one of the 15 AI configuration-pair means.

Configuration	Summary	Strengths	Weak.	Quest.
G-ARev	0.856	0.852	0.841	0.742
G-DRev	0.812	0.804	0.790	0.688
G-RGrd	0.895	0.837	0.826	0.744
D-ARev	0.817	0.864	0.850	0.764
D-DRev	0.832	0.800	0.795	0.676
D-RGrd	0.918	0.818	0.796	0.707

Table 10: AI-human centroid similarity by configuration and section. Each value compares an AI section with the centroid of the human reviews for the same paper.

Section	Same F.	Same M.	Cross	Δ
Summary	0.891	0.825	0.817	+0.066
Strengths	0.856	0.839	0.820	+0.016
Weaknesses	0.857	0.835	0.822	+0.021
Questions	0.832	0.767	0.758	+0.065

Table 11: Framework-versus-model decomposition. Same F. denotes same framework across models; Same M. denotes same model across frameworks; Cross denotes different frameworks and different models. Δ is Same F. $-$ Same M.

paper p, let P_p denote the set of chunks from the target paper. For a review unit r_i, we define the target top-k support as:

$$\mathrm{TargetTopK}_i = \frac{1}{k} \sum_{c \in \mathrm{TopK}_k(P_p; r_i)} \cos\big(f(r_i), f(c)\big), \tag{7}$$

where $\mathrm{TopK}_k(P_p; r_i)$ returns the k chunks in the target paper with the highest cosine similarity to r_i.

We then construct B balanced hard-negative pools, denoted as $H_{p,b}^{\mathrm{bal}}$, from semantically similar non-target papers. For each pool, we compute the average similarity between the review unit and its top-k negative chunks:

$$\mathrm{NegTopK}_i = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{k} \sum_{c \in \mathrm{TopK}_k(H_{p,b}^{\mathrm{bal}}; r_i)} \cos\big(f(r_i), f(c)\big). \tag{8}$$

The final support margin is:

$$\mathrm{SupportMargin}_i = \mathrm{TargetTopK}_i - \mathrm{NegTopK}_i. \tag{9}$$

At the group level, we report the mean support margin across review units:

$$\mathrm{SupportMargin}(p, g) = \frac{1}{|R_{p,g}|} \sum_{r_i \in R_{p,g}} \mathrm{SupportMargin}_i. \tag{10}$$

In our main setting, we use $k = 3$ and $B = 20$. Hard-negative papers are sampled from the 10 non-target papers closest to the target paper under paper-level embedding similarity.

Interpretation. A higher Balanced Top-k Evidence Support Margin indicates stronger paper-specific semantic support: the critique remains more anchored in the submitted paper than in generally similar papers.

E.2 Residual Semantic Dispersion beyond Local Evidence

To measure meta-level evaluative synthesis beyond local evidence, we compute residual semantic dispersion after subtracting the nearest same-paper chunk. This residual axis captures the transformation layer between evidence localization and final judgment.

For each review unit r_i, we first retrieve the nearest chunk from the target paper:

$$c_i^+ = \arg\max_{c \in P_p} \cos\big(f(r_i), f(c)\big). \tag{11}$$

Let

$$x_i = f(r_i), \qquad y_i = f(c_i^+). \tag{12}$$

The residual vector is then defined as:

$$u_i = x_i - y_i. \tag{13}$$

The residual trajectory for paper p and review group g is:

$$U_{p,g} = \{u_1, u_2, \ldots, u_n\}. \tag{14}$$

We compute four residual statistics.

First, the mean residual norm:

$$\mathrm{ResNorm}(U) = \frac{1}{n} \sum_{i=1}^{n} \|u_i\|_2. \tag{15}$$

Second, residual effective rank. Let λ_j be the eigenvalues of the covariance matrix of the centered residual trajectory, and define:

$$q_j = \frac{\lambda_j}{\sum_\ell \lambda_\ell}. \tag{16}$$

Then:

$$D_{\mathrm{eff}}(U) = \exp\left(-\sum_j q_j \log q_j \right). \tag{17}$$

Third, residual information volume per unit:

$$V_{\mathrm{unit}}(U) = \frac{1}{n} \cdot \frac{1}{2} \log\det\left(I + \alpha Z Z^\top\right), \tag{18}$$

where Z is the centered residual trajectory matrix.

Fourth, normalized residual information volume. Let

$$\widehat{u}_i = \frac{u_i}{\max(\|u_i\|_2, \epsilon)}, \tag{19}$$

and let $\widehat{Z}$ be the centered matrix formed by the normalized residual vectors $\{\widehat{u}_i\}$. We define:

$$\widetilde{V}_{\mathrm{unit}}(U) = \frac{1}{n} \cdot \frac{1}{2} \log\det\left(I + \alpha \widehat{Z} \widehat{Z}^\top\right). \tag{20}$$

Interpretation. Residual semantic dispersion measures post-local-evidence semantic variation: how much the review moves from local textual anchoring toward evaluative judgment. Higher residual dispersion can reflect cross-fragment integration, field-level standards, or severity calibration.

E.3 Support-Scaled Residual Volume for Human-AI Review

Section 5.5 uses Support-Scaled Residual Volume (SSRV) to analyze whether Human+AI reviews concentrate evaluative content into evidence-supported directions. This metric combines the two axes defined above: the support margin M_i, which measures whether a critique unit is anchored in the target paper, and the evidence-conditioned residual u_i, which captures evaluative movement beyond nearest local evidence.

For each review unit r_i, we first keep only positive paper-specific support:

$$m_i^+ = \max(M_i, 0). \tag{21}$$

Let the mean positive support for paper p and review group g be

$$\overline{M^+}_{p,g} = \frac{1}{n}\sum_{i=1}^{n} m_i^+. \tag{22}$$

We then define normalized support weights

$$\omega_i = \frac{m_i^+}{\max(\overline{M^+}_{p,g}, \epsilon)}, \tag{23}$$

where ϵ is a small constant used only to avoid division by zero. These weights emphasize residual directions attached to paper-specific evidence while down-weighting unsupported residual movement.

Given residual vectors u_i, we compute the weighted residual mean

$$\mu_w = \frac{\sum_i \omega_i u_i}{\sum_i \omega_i}, \tag{24}$$

when $\sum_i \omega_i > 0$; otherwise, the weighted residual matrix is set to zero. We form support-weighted centered residual rows:

$$\widetilde{u}_i^{(w)} = \sqrt{\omega_i}(u_i - \mu_w), \tag{25}$$

and let $\widetilde{U}_{p,g}^{(w)}$ denote the matrix containing these rows.

We define grounded weighted residual volume as

$$GWV_{p,g}^{unit} = \frac{1}{2n}\log\det\left(I + \alpha \widetilde{U}_{p,g}^{(w)} \widetilde{U}_{p,g}^{(w)\top}\right), \tag{26}$$

where α is the same scaling constant used for residual information volume. GWV^{unit} captures the volume of residual evaluative movement after emphasizing units with positive paper-specific support.

Finally, we scale this weighted residual volume by the mean positive support:

$$SSRV_{p,g} = GWV_{p,g}^{unit} \cdot \overline{M^+}_{p,g}. \tag{27}$$

Thus, SSRV is high only when a review contains residual evaluative structure and that structure is broadly supported by target-paper evidence. Unsupported semantic expansion alone does not increase SSRV.

We also report grounded weighted effective dimension, GWD_{eff}, to describe how dispersed the supported residual content is. Let λ_j be the eigenvalues of the covariance matrix of $\widetilde{U}_{p,g}^{(w)}$, and define

$$q_j = \frac{\lambda_j}{\sum_\ell \lambda_\ell}. \tag{28}$$

Then

$$GWD_{eff} = \exp\left(-\sum_j q_j \log q_j\right). \tag{29}$$

A larger GWD_{eff} means that supported residual content is spread across more semantic directions, while a smaller value means that it is concentrated into fewer directions. We do not interpret low GWD_{eff} as a standalone quality signal. Instead, we interpret it jointly with SSRV: high SSRV with lower GWD_{eff} indicates compact evidence-supported synthesis, where substantial supported evaluative content is concentrated into fewer critique directions.

E.4 Sampling and Matching

To reduce the influence of review length on embedding geometry, we apply per-paper capped sampling. All AI–human comparisons are computed at the per-paper level:

$$\Delta M_p = M(X_{p,\mathrm{AI}}) - M(X_{p,\mathrm{Human}}), \tag{30}$$

and then aggregated across papers with bootstrap confidence intervals.

E.5 Complete Results

Table 12 reports the full Balanced Top-k Evidence Support Margin ranking under both W+Q and weaknesses-only settings.

Setting	System	Evidence Support Margin	Target Top-k	Negative Top-k
W+Q	GPT-5.5 ReviewGrounder	0.1125	0.6742	0.5616
W+Q	DeepSeek ReviewGrounder	0.1089	0.6744	0.5655
W+Q	DeepSeek AgentReview	0.0873	0.6461	0.5588
W+Q	GPT-5.5 AgentReview	0.0798	0.6322	0.5524
W+Q	Human	0.0759	0.6146	0.5387
W+Q	GPT-5.5 DeepReviewer	0.0700	0.6100	0.5399
W+Q	DeepSeek DeepReviewer	0.0670	0.5971	0.5301
Weaknesses-only	GPT-5.5 ReviewGrounder	0.1181	0.6885	0.5705
Weaknesses-only	DeepSeek ReviewGrounder	0.1051	0.6760	0.5709
Weaknesses-only	GPT-5.5 DeepReviewer	0.1005	0.6587	0.5582
Weaknesses-only	DeepSeek AgentReview	0.0843	0.6346	0.5503
Weaknesses-only	DeepSeek DeepReviewer	0.0813	0.6178	0.5366
Weaknesses-only	GPT-5.5 AgentReview	0.0810	0.6254	0.5444
Weaknesses-only	Human	0.0759	0.6146	0.5386

Table 12: Complete Balanced Top-k Evidence Support Margin results. Higher values indicate stronger paper-specific evidence support relative to balanced hard-negative evidence pools.

The results show that ReviewGrounder consistently achieves the highest evidence support. Human review is not evidence-free: under the balanced metric, it occupies a moderate support region rather than collapsing to zero or negative support.

Residual Semantic Dispersion. Table 13 reports the full residual semantic dispersion ranking under both W+Q and weaknesses-only settings. Human review achieves the highest composite residual dispersion score in both settings. This supports the interpretation that human reviews preserve more post-local-evidence semantic movement, consistent with meta-level evaluative synthesis beyond nearest-chunk textual anchoring.

Additional Diagnostic Maps. The main text uses the Balanced Top-k Evidence Support–Residual Semantic Dispersion map as the primary Figure 3. We include additional maps here to show that the separation between local evidence grounding and meta-level evaluative synthesis is not an artifact of one specific projection.

Summary. The appendix results support the main conclusion of research question 4: human and AI reviews differ systematically in how they balance evidence grounding and residual evaluative content. ReviewGrounder is strongest in paper-specific evidence support but more semantically constrained. Human review is moderately grounded but retains higher residual semantic dispersion beyond local evidence. DeepReviewer is more report-like and section-sensitive, showing higher residual variation in W+Q but improved grounding in weaknesses-only.

This supports a role-based view of AI-assisted reviewing: AI systems are well positioned to strengthen evidence grounding, while human reviewers remain central to evaluative synthesis and final judgment.

F Human-AI Review Study Materials

This appendix reports the materials used for the human-led review study in Section 5.1. The main text frames the study as a responsibility-allocation test; this section provides the protocol details needed to audit the assisted condition.

F.1 Review Copilot Workflow

Review Copilot was used as the study environment. The workspace combines PDF reading, evidence-grounded question answering, draft generation, structured point editing, and final review generation. Reviewers can inspect the submitted paper, search within the PDF, ask questions grounded in paper content, follow evidence links back to the source text, and edit summary, strength, weakness, and question points before finalizing the review.

The workflow prevents the initial AI draft from becoming a final review by default. An initial draft can give reviewers a starting structure, but reviewers must inspect the paper through question-answering interactions, revise the draft, and decide which critique points to retain before producing the final review. Point-level editing and locking allow reviewers to preserve their own judgments during final generation.

Consistent with the rating-calibration analysis in Section 4, **Review Copilot does not generate an official score or recommendation;** consistent with the diversity analysis, AI-generated critiques are not counted as additional panel voices.

Setting	System	Score	Mean Residual Norm	Residual D_{eff}	Residual V_{unit}
W+Q	Human	70.8333	0.8445	9.3679	2.4048
W+Q	GPT-5.5 DeepReviewer	58.3333	0.8515	9.6469	2.3862
W+Q	DeepSeek DeepReviewer	50.0000	0.8641	9.3284	2.3934
W+Q	DeepSeek AgentReview	45.8333	0.8052	9.7617	2.3409
W+Q	GPT-5.5 AgentReview	45.8333	0.8240	9.7392	2.3557
W+Q	GPT-5.5 ReviewGrounder	41.6667	0.7695	9.8233	2.3187
W+Q	DeepSeek ReviewGrounder	37.5000	0.7698	8.6949	2.3429
Weaknesses-only	Human	75.0000	0.8445	9.3679	2.4048
Weaknesses-only	GPT-5.5 DeepReviewer	58.3333	0.7950	5.7620	2.4144
Weaknesses-only	DeepSeek DeepReviewer	50.0000	0.8394	7.6201	2.3921
Weaknesses-only	GPT-5.5 ReviewGrounder	45.8333	0.7507	7.8299	2.3675
Weaknesses-only	DeepSeek AgentReview	41.6667	0.8179	9.7378	2.3426
Weaknesses-only	GPT-5.5 AgentReview	41.6667	0.8311	9.6900	2.3512
Weaknesses-only	DeepSeek ReviewGrounder	37.5000	0.7673	4.2545	2.3959

Table 13: Complete residual semantic dispersion results. The score is a composite residual-dispersion ranking; higher values indicate more semantic variation after subtracting the nearest same-paper evidence chunk.

F.2 Reviewer Instructions

Figure 12 reproduces the short guide given to reviewers before they used Review Copilot. The guide defines the system as optional assistance, states that AI output is not a completed review, and specifies the sequence reviewers should follow before final export. It makes the assisted condition human-led in practice by requiring reviewers to inspect the paper, revise the structured draft, and decide which points enter the final report.

F.3 Interaction Logs

The study records how reviewers handle AI-generated artifacts before they enter the final report. For each displayed artifact, the log stores anonymized reviewer and paper identifiers, artifact type, suggestion identifier, reviewer action, edit status, and rejection reason when available. The resulting fields are *reviewer_id*, *paper_id*, *artifact_type*, *suggestion_id*, *action*, *edited*, and *rejection_reason*. These logs support the artifact mediation analysis in Section 5.4.

F.4 Recruitment And Payment

Participants were invited because they had prior reviewing experience at top AI conferences. Participation was voluntary. No monetary compensation was provided. As a non-monetary benefit, participants will receive access to improved versions of Review Copilot developed from the study feedback, but this access was not tied to any required positive evaluation of the system.

G Statement on the Use of AI Assistants

AI assistants were used during the writing and research process in a limited supporting role. They helped identify and correct grammar and wording issues, improve clarity of presentation, and provide suggestions for organizing and interpreting the data analysis. All substantive research decisions, experimental design choices, data processing, result interpretation, and final manuscript content were reviewed and finalized by the authors.

Professionalism Prompt

```
You are an expert peer-review evaluation judge.

Evaluate professionalism as civility and non-abusive peer-review tone only.

- Evaluate the review as one review-level artifact using only the four provided sections: Summary, Strengths, Weaknesses,
      and Questions.
- Judge whether the review keeps a respectful, formal, non-hostile academic tone toward the authors, the work, and any
     people or groups mentioned.
- Do not judge factual correctness, paper grounding, rating correctness, decision correctness, review completeness,
     usefulness, actionability, relevance, or formatting quality.
- Ignore non-review content, system artifacts, irrelevant content, weak organization, missing sections, grammar mistakes,
      or incoherence unless the text itself contains abusive, discriminatory, mocking,
hostile, or emotionally inappropriate language.
- Strong technical criticism is professional when it targets the work, evidence, claims, experiments, or writing rather
     than attacking people.
- Mark severity="major" for personal attacks; insults; sarcasm or mockery; hostile, contemptuous, or emotionally charged
     language; discriminatory or identity-based remarks; accusations about author
competence, integrity, motives, or effort; or clearly abusive/unsafe language.
- Mark severity="minor" for mildly unprofessional tone such as somewhat dismissive wording, unnecessarily harsh phrasing,
      excessive frustration, informal snark, or emotionally loaded criticism, when it does
not rise to direct attack or hostility.
- Mark severity="none" for a normal respectful academic review, including blunt but evidence-focused criticism.
- is_professional must be true for severity "none" or "minor", and false for "major".

Return JSON only, with this exact shape:
{
  "is_professional": true,
  "severity": "none" | "minor" | "major",
  "violations": [
    {
      "type": "personal_attack | insult | mockery_or_sarcasm | hostile_tone | emotional_language |
     discriminatory_language | author_motive_attack | unsafe_claim | other",
      "span": "...",
      "reason": "..."
    }
  ],
  "reason": "..."
}
```

Figure 6: Exact LLM-as-judge prompt for the Professionalism metric.

Unsupported Prompt

```
You are an expert peer-review grounding judge.

Evaluate whether each fixed feedback unit is supported by the provided paper evidence.

Definitions:
- supported: the unit's main concrete factual/descriptive claim about the paper is supported by the evidence.
- partially_supported: some factual content is supported, but the unit has a clear overstatement, missing qualifier, or
      unsupported subclaim.
- unsupported: the unit makes a concrete factual/descriptive claim about the paper, but the provided evidence does not
      support it.
- contradicted: the unit is clearly inconsistent with or contradicted by the provided evidence.
- not_verifiable: the unit is mainly subjective opinion, generic advice, preference, or value judgment without a concrete
       paper-factual claim.

Rules:
- The input contains a fixed list of feedback_units from one review section.
- Do not split, merge, drop, reorder, or rewrite the feedback_units.
- Return exactly one judgment for each input unit, preserving unit_index and unit_hash.
- Use only the provided evidence_chunks from the paper. Do not use outside knowledge.
- Each unit also includes candidate_evidence_chunk_ids retrieved for that unit. Prefer those chunks, but you may cite any
      provided evidence_chunk if it supports your judgment.
- Judge paper-grounding only. Do not judge whether the review is useful, polite, complete, or correct as an opinion.
- Strong criticism can be supported if the evidence shows the relevant issue.
- Do not penalize a unit merely because it is negative.
- If a concrete claim requires evidence and the evidence is absent, label unsupported.
- If the evidence partly supports the claim but the unit exaggerates or adds unsupported specifics, label
      partially_supported.
- If the unit only says something broad like "the method is interesting", "the writing could be improved", or "more
      experiments would help" without a specific factual claim, label not_verifiable.
- evidence_chunk_ids must contain only ids from the provided evidence_chunks.
- Do not echo unit_text in the output.

Return JSON only, with this exact compact shape.
Do not include claim_summary, reason, explanations, unit_text, or any other fields:
{
  "feedback_units": [
    {
      "unit_index": 1,
      "unit_hash": "...",
      "label": "supported | partially_supported | unsupported | contradicted | not_verifiable",
      "is_verifiable": true,
      "unsupported_error_type": "missing_evidence | overstatement | wrong_attribution | contradicted_by_evidence |
     unsupported_comparison | other | null",
      "evidence_chunk_ids": ["para_0001"]
    }
  ]
}
```

Figure 7: Exact LLM-as-judge prompt for the Unsupported metric.

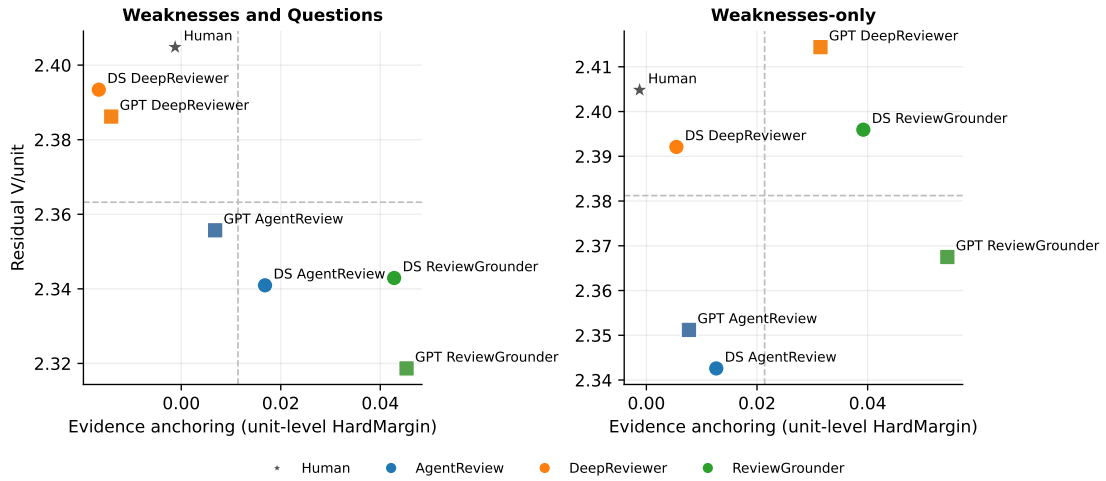

Figure 8: Unit-level HardMargin Evidence Anchoring versus residual semantic dispersion. The x-axis measures strict hard-negative evidence anchoring, while the y-axis measures residual semantic variation beyond local evidence. This map shows that strong local anchoring and meta-level evaluative synthesis are separable dimensions.

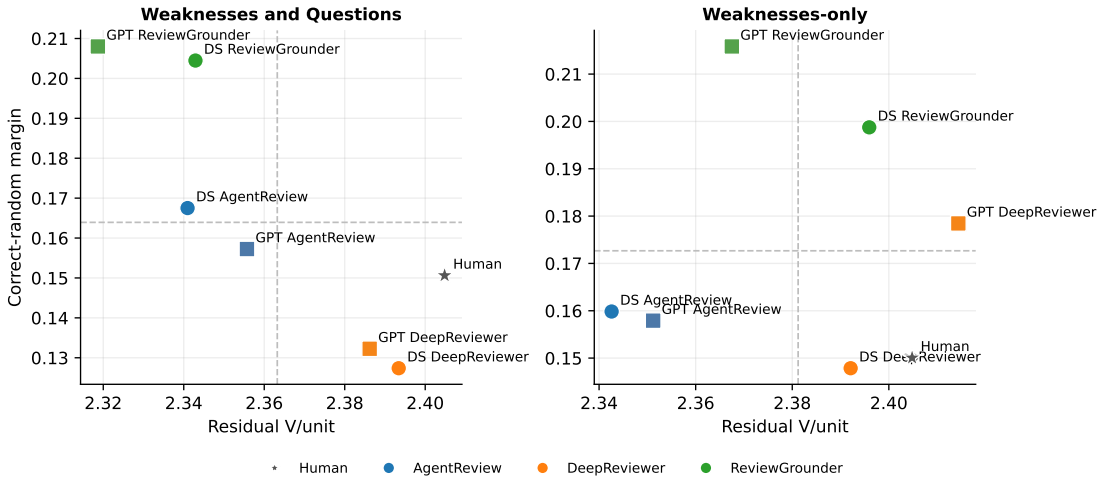

Figure 9: Correct-random evidence margin versus residual semantic dispersion. This view further confirms that paper-specific textual anchoring and post-evidence evaluative transformation are not equivalent.

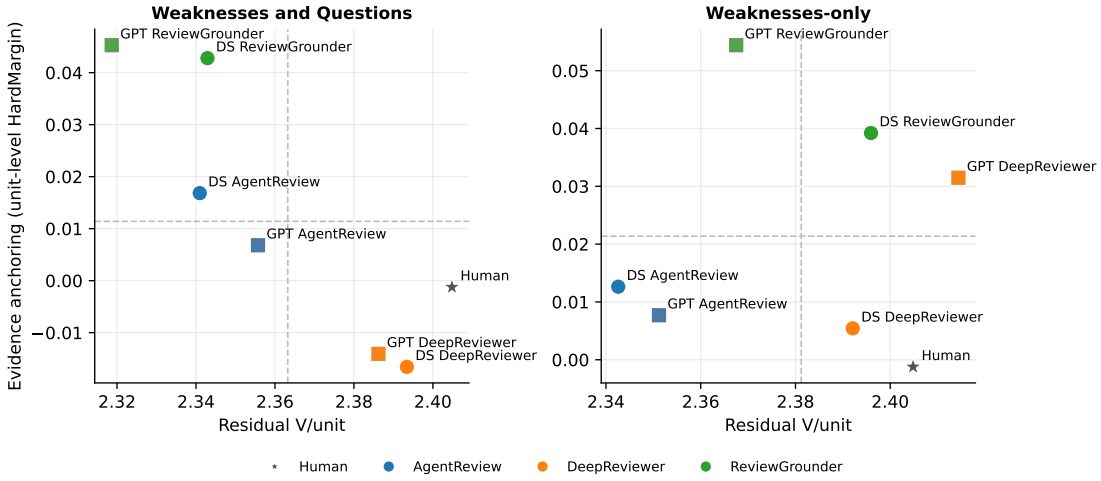

Figure 10: Residual information volume per unit versus unit-level HardMargin Evidence Anchoring. This projection shows the same separation between strict evidence anchoring and residual evaluative synthesis.

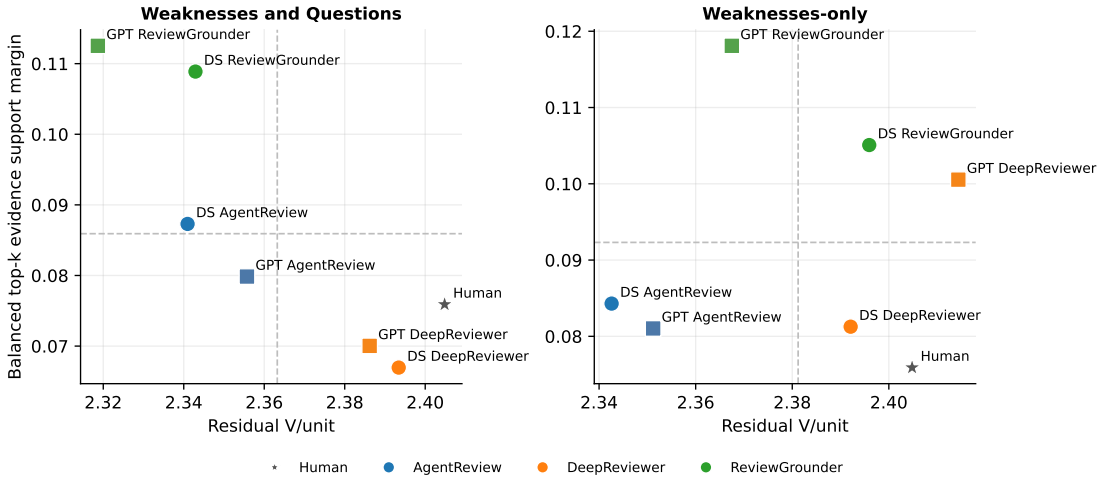

Figure 11: Residual information volume per unit versus Balanced Top-k Evidence Support Margin. This map provides an alternative view of the main Figure 3.

Reviewer Guide Used in the Human-Led Review Study

```
## Role of the system.
Treat Review Copilot as optional assistance for reading, checking, and drafting.
AI output is not a completed review. The final critique, score, and
recommendation remain the reviewer's responsibility.

## Stage 1: initial draft.
When starting a new paper, generate an initial draft based only on the submitted
paper. Use this draft as a reading scaffold and a source of candidate issues, not
 as a final judgment.

## Stage 2: required paper-grounded QA.
Before the final export step becomes available, complete at least seven question-
answering interactions about the paper. Use these interactions to inspect claims,
 check details, and follow evidence links back to the source PDF. The system does
 not allow reviewers to skip directly from the initial draft to the final review.

## Stage 3: edit and lock review points.
Return to the structured draft and revise it point by point. You may add, delete,
 rewrite, or keep individual summary, strength, weakness, and question items.
Lock points that reflect your own judgment or that you want preserved during
final generation. Locked content is kept fixed when the final review is generated
.

## Stage 4: final review generation.
After the QA threshold and editing steps are complete, generate the final review.
 The final generation summarizes the reviewer's inspected evidence, edited points
, and locked judgments. Reviewers should make any final corrections before
submission.
```

Figure 12: Participant-facing reviewer guide used in the human-led review study.